Kwaliteit van Nederlandse Spraaksynthese

Lam, D.H.W.
4298772
Begeleider: G. Bloothooft
Tweede beoordelaar: K.E. Overvliet
Bacheloreindwerkstuk 7.5 ECT

Bachelor Kunstmatige Intelligentie
Faculteit Geesteswetenschap
Universiteit Utrecht
Nederland
30 april 2018

Samenvatting

In deze scriptie wordt een onderzoek besproken, dat uitgevoerd is naar de gemiddelde beoordeling van Nederlandse spraaksynthese van Text-To-Speech systemen. Dit onderzoek is uitgevoerd, omdat er geen duidelijk beeld was van de beoordeling van Nederlandstalige Text-To-Speech-systemen door eindgebruikers.

Er zijn 8 systemen van o.a. de bedrijven Nuance, Fluency, Ivona, Acapela, Readspeaker en ResponsiveVoice onderzocht. De resultaten zijn door middel van een Mean Opinion Score test in een vorm van een Enquête afgenomen, ingevuld door studenten uit het hoger onderwijs, verdeeld over twee groepen.

Uit de resultaten is gebleken dat de kwaliteit van Nederlandse spraaksynthese als geheel nog niet wordt beoordeeld op het niveau van menselijke spraak, maar dat er wel verschillen bestaan tussen verschillende Text-To-Speech systemen.

Inhoudsopgave

1 Introductie

1.1 Spraaksynthese

Wie een beetje bekend is met beroemdheden, dan zal de naam Stephen Hawking mogelijk bekend klinken. Stephen Hawking was een Britse wetenschapper die onder andere bijdrage leverde over de theorie van zwarte gaten, maar velen zullen hem eerder herkennen door zijn manier van praten. In 1962 werd zijn spierziekte ALS geconstateerd en jaren later verloor hij zijn stem om te kunnen praten. De rolstoel van Stephen Hawking is uitgerust met een spraakgenererende apparaat om voor hem te spreken. Op zijn computerscherm is een cursor aanwezig waar hij met bewegingen van zijn wang kon besturen. Zo kon hij na het selecteren van beginletters, woorden kiezen om dan zinnen te vormen die hij dan weer stuurt naar een speech synthesizer om spraak te genereren. Dit was het iconische mechanische stem die men tegenwoordig kent (Hawking, 2018) [1]. Deze techniek die de bekende wetenschapper gebruikte wordt ook wel spraaksynthese genoemd. Spraaksynthese is een proces waarbij een computer menselijke spraak op kunstmatige wijze produceert en het komt steeds vaker voor in de maatschappij. Deze techniek wordt gebruikt in GPS-systemen, smartphones en de zogeheten smart speakers zoals Google Home of Amazon Echo die recent op de markt zijn gekomen, maar het wordt normaliter gebruikt om tekst voor te lezen in een vorm van een text-to-speech (TTS) systeem. Bij het Centraal Examen kunnen leerlingen met dyslexie of andere cognitieve beperkingen gebruik maken van text-to-speech-software om een examen voor te lezen en websites zoals thuisarts.nl of de officiële website van gemeente Utrecht hebben tegenwoordig een voorleesknop om de inhoud van de website voor te lezen.

Op het internet of mobiele apparaten zoals tablets of laptops kan men vaak extensies of applicaties met spraaksynthese-software downloaden. Vaak klinken deze kunstmatige stemmen die gratis worden aangeboden erg mechanisch en houterig, maar er bestaan ook kwalitatief betere stemmen ontwikkeld door wetenschappers of bedrijven die mensen met visuele of communicatieve beperkingen kunnen helpen. In Ruiter (2010) [2] wordt een onderzoek besproken naar welke digitale basisvoorzieningen voorrang moesten verkrijgen voor verdere onderzoek en ontwikkeling voor gebruikers van de Nederlandse taal met communicatieve beperkingen. In de conclusie werd spraaksynthese als één van deze digitale basisvoorzieningen benoemd. Als deze techniek zou worden verbeteren, zou er bijvoorbeeld betere voorlees-software kunnen komen of betere synthetische stemmen die voor mensen met spraakproblemen kunnen spreken. De kwaliteit van de digitale basisvoorziening zou dan ook worden verbeterd voor de mensen met communicatieve of visuele beperkingen.

1.2 Probleemstelling & Onderzoeksdoel

1.2.1 Achtergrond probleemstelling

Momenteel is er geen duidelijk beeld van de beoordeling van Nederlandstalige Text-To-Speech-systemen door eindgebruikers. Er ontbreekt ook een beoordeling van de individuele aspecten van TTS-systemen. Dit maakt het niet alleen voor wetenschappers en onderzoekers, die continu werken aan het verbeteren van TTS-systemen, maar ook bedrijven gespecialiseerd in het produceren van TTS-systemen lastig om in te schatten welke zaken het meest kritisch zijn bij het voor eindgebruikers beter maken van dit soort systemen. Er werden 8 systemen van o.a. de bedrijven Nuance, Fluency, Ivona, Acapela, Readspeaker en ResponsiveVoice gekozen om te onderzoeken.

1.2.2 Onderzoeksdoel

Het doel van het onderzoek, dat in deze scriptie besproken wordt, is het verkrijgen van een beter beeld van de beoordeling van Nederlandstalige TTS systemen. Belangrijk daarbij is het opsplitsen van deze beoordelingen in meerdere aspecten, zodat het inzichtelijker wordt welke zaken belangrijke verbeterpunten zijn bij toekomstige ontwikkelingen aan (Nederlandstalige) TTS systemen.

1.2.2.1 Onderzoeksvragen

Aansluitend op het doel van dit onderzoek, is de hoofdvraag van dit onderzoek als volgt geformuleerd:

Wat is de huidige kwaliteit van Nederlandse spraaksynthese?

Om de hoofdvraag te kunnen beantwoorden, zijn de volgende deelvragen gesteld:

1) Hoe wordt de natuurlijkheid en verstaanbaarheid van spraak beoordeeld?

2) Worden alle stemmen even goed beoordeeld?

3) Wat zijn de belangrijkste verbeterpunten van Nederlandstalige Text-To-Speech?

4) Kan Nederlandstalige Text-To-Speech voor meer toepassingen worden ingezet als er verbeteringen worden doorgevoerd?

Deze deelvragen hebben betrekking op de kwaliteit op Nederlandstalige spraaksynthese. De verstaanbaarheid van spraak is één van de belangrijkste factoren voor TTS systemen om mogelijk ingezet te worden voor verschillende doeleindes. TTS systemen zullen hoogstwaarschijnlijk niet gebruikt worden wanneer deze niet verstaan wordt. De natuurlijkheid kan invloed hebben voor het gebruik van een TTS systeem.

In Nederland zijn verschillende TTS systemen van verschillende bedrijven beschikbaar. Met de tweede deelvraag wordt er gekeken of alle TTS systemen dezelfde kwaliteit beschikken of dat er toch verschillen blijken te zijn.

Bij het beoordelen op de kwaliteit van TTS systemen worden hoogstwaarschijnlijk aspecten genoemd waar een systeem niet goed presteert. Met de derde deelvraag kan worden aangewezen op welke aspect of aspecten Nederlandstalige Text-To-Speech verbeterd moet worden om de kwaliteit ook te verbeteren.

Als laatst wordt de vraag gesteld waar TTS systemen nog meer toegepast kan worden.

1.3 Structuur scriptie

In sectie 2 wordt algemeen beschreven hoe een TTS systeem werkt en hoe deze worden geëvalueerd, in sectie 3 wordt beschreven hoe het onderzoek is uitgevoerd met welke middelen, in sectie 4 worden de resultaten van het uitgevoerde onderzoek beschreven en in sectie 5 worden de onderzoeksvraag en deelvragen beantwoord en volgt er een discussie.

2 Theoretische Achtergrond

2.1 Text-To-Speech systeem

2.1.1 Wat is een Text-To-Speech systeem?

Een text-to-speech systeem moet vanuit een tekst, bestaande uit letters en leestekens, spraak genereren. Een TTS systeem bestaat altijd uit twee hoofdcomponenten waar verschillende taken worden uitgevoerd. We noemen deze twee componenten de taalkundige en spraakgenererende component (Rietveld, 2016) [3].

2.1.2 Taalkundige component

De taalkundig component zorgt ervoor dat de tekst wordt omgezet naar een fonetische representatie, zodat de spraakgenererende component weet hoe een tekst uitgesproken moet worden. Het bewerken van de tekst wordt in twee fasen gedaan.

De tekst moet eerst ontleed worden naar korte uitspreekbare zinnen. Dit kan gedaan worden met een proces genaamd *tokenization*. Tokenization houdt in dat een tekst opgedeeld wordt in zinnen of woorden na spaties of interpunctie. De gesplitste lexicale eenheden worden tokens genoemd. Voor een TTS systeem worden teksten voornamelijk gesplitst in zinnen, maar *tokenization* voor zinnen verloopt niet altijd makkelijk. Een zin hoeft niet altijd te eindigen op bijvoorbeeld een punt. Dit kan namelijk ook andere leestekens zijn die bijzinnen inleiden. De meeste tokenization algoritmes voor zinnen zijn met machine learning getraind. In Jurasfky en Martin (2008) [4] wordt er een voorbeeld gegeven van een tokenization algoritme. De classificeerder werd getraind om te voorspellen of een bepaalde token een grens is van een zin. Hierbij krijgt de classificeerder een training set waar de tokens die de grenzen van zinnen al bepalen met de hand zijn gelabeld. De bedoeling is dat de classificeerder uiteindelijk kan voorspellen of een desbetreffende token T, de grens is van een zin. Dit kan gedaan worden door bijvoorbeeld te kijken wat de kans is dat token T op het einde van een zin staat en wat de kans is dat de volgende token na token T de eerste woord is van een zin, wat we ook **features** noemen. Het aantal features en de grootte van de training sets bepalen hoe goed dit algoritme presteert.

Vervolgens moeten de woorden in de zinnen geclassificeerd worden in groepen waaraan regels voor uitspraak toegekend kunnen worden. Er bestaan ook niet-standaard woorden zoals nummers, acroniemen of afkortingen, welke ook moeten worden uitgesproken door een TTS systeem. Hiervoor moet er een algoritme komen, welke bepaalt hoe deze niet-standaard woorden uitgeschreven worden, zodat de juiste fonetische uitspraak toegekend kan worden. Ook hiervoor kan er een classificeerder worden gebruikt die op dezelfde wijze wordt getraind als de tokenization algoritme van Jurasfky en Martin [4]. Er wordt eerst gekeken of een betreffende niet-standaard woord bestaat uit alfabetische schrift, numerieke tekens of anders. Vervolgens wordt het woord geclassificeerd met een label. De labels bepalen hoe de niet-standaard woorden uitgeschreven op schrift staan. Zo worden bijvoorbeeld telefoonnummers als aparte getallen uitgesproken en de afkorting N.Y. wordt volledig uitgesproken als New York. In Sproat et al. (2001) [5] worden meerdere technieken voor dit proces besproken. In figuur 2.1 staat een overzicht aan labels die gebruikt worden.

	EXPN	abbreviation	*adv, N.Y, mph, gov't*
alpha	LSEQ	letter sequence	*CIA, D.C, CDs*
	ASWD	read as word	*CAT*, proper names
	MSPL	misspelling	*geogaphy*
	NUM	number (cardinal)	*12, 45, 1/2, 0·6*
	NORD	number (ordinal)	*May 7, 3rd, Bill Gates III*
	NTEL	telephone (or part of)	*212 555-4523*
	NDIG	number as digits	*Room 101*
N	NIDE	identifier	*747, 386, I5, pc110, 3A*
U	NADDR	number as street address	*5000 Pennsylvania, 4523 Forbes*
M	NZIP	zip code or PO Box	*91020*
B	NTIME	a (compound) time	*3·20, 11:45*
E	NDATE	a (compound) date	*2/2/99, 14/03/87* (or US) *03/14/87*
R	NYER	year(s)	*1998, 80s, 1900s, 2003*
S	MONEY	money (US or other)	*$3·45, HK$300, Y20,000, $200K*
	BMONEY	money tr/m/billions	*$3·45 billion*
	PRCT	percentage	*75%, 3·4%*
	SPLT	mixed or "split"	*WS99, x220, 2-car* (see also SLNT and PUNC examples)
	SLNT	not spoken,	word boundary or emphasis character:
M		word boundary	*M.bath, KENT*RLTY, _really_*
I	PUNC	not spoken,	non-standard punctuation: "***" in
S		phrase boundary	*$99,9K***Whites*, "..." in *DECIDE... Year*
C	FNSP	funny spelling	*sllooooowww, sh*t*
	URL	url, pathname or email	*http://apj.co.uk, /usr/local, phj@tpt.com*
	NONE	should be ignored	ascii art, formatting junk

Figuur 2.1: Overzicht van non-standaard woorden labels

2.1.2.1 Fonetische analyse

Vervolgens wordt er een fonetische codering gedaan. De uitspraakcodering in fonetische schrift worden gekoppeld aan tekstwoorden. Hiervoor kan het proces grafeem-naar-foneem gebruikt worden, afgekort als G2P van de Engelse term grapheme-to-phoneme. Een grafeem is een letter en een foneem noemen we een klankvormige eenheid zoals een medeklinker of klinker. De methode G2P zoekt aan de hand van de symbolische representatie de juiste uitspraak die bij het symbool hoort. Er bestaan meerdere varianten van G2P.

Een simpele techniek van G2P is met behulp van een uitspraakwoordenboek. De uitspraken van grafemen worden opgezocht in een uitspraakwoordenboek. Sommige tekstwoorden staan qua uitspraak volledig opgeslagen in dit uitspraakwoordenboek, zodat deze direct uit gehaald kan worden. Woorden die er niet als geheel staan, worden zo mogelijk weer opgedeeld in kleinere stukken die wel in het uitspraakwoordenboek staan. Het nadeel van deze aanpak hiervan is dat het maken van een uitspraakwoordenboek voor een TTS systeem veel tijd en geheugen vraagt. Een uitspraakwoordenboek is daardoor immers ook beperkt, terwijl er van een TTS systeem verwacht wordt dat deze elk willekeurig woord kan uitspreken (Dutoit, 1997) [6].

Een ander variant van G2P is rule-based G2P. Een rule-based G2P gebruikt vaste regels hoe een grafeem uitgesproken moet worden. Elke letter wordt bijvoorbeeld gekoppeld aan bijbehorende fonemen die bij deze letter zouden kunnen behoren. Hierdoor hoeft er geen groot uitspraakwoordenboek gebruikt te worden en wordt elke willekeurige woord qua uitspraak ondervangen. In het algemeen presteert deze rule-based G2P goed, maar ook hier zijn nadelen aan verbonden. Het maken van deze regels is zeer lastig, aangezien natuurlijke taal vaak uitzonderingsgevallen bevat. Als voorbeeld in het Nederlands heb je bijvoegelijke naamwoorden die eindigen met -lijk. De **ij** in woorden als *eerlijk* of *vrolijk* worden uitgesproken met een stomme e, maar in woorden zoals *gelijk* wordt de **ij** met een normale ij-klank uitgesproken. In het Engels wordt bijvoorbeeld de **eo** in *people*, *leopard* en *leotard* anders uitgesproken. Deze uitzonderingsgevallen moeten eigen regels hebben voor de rule-based G2P methode wat mogelijk veel tijd kan kosten. Vaak wordt er dan ook gebruik gemaakt van een woordenboek om deze uitzonderingsgevallen op te vangen.

Een andere versie van G2P is data-driven G2P. Deze methode van G2P wordt met behulp van machine learning gebouwd. Deze methode wordt aan de hand van voorbeelden van grafeem en uitspraak getraind. Hierdoor zal de data-driven G2P een patroon voor het

omzetten van grafeem naar foneem herkennen. Aan de hand hiervan zal een data-driven G2P leren te voorspellen welke fonemen het best bij de grafemen behoren. Wat hier nog wel eens fout kan gaan is dat de kwaliteit van de trainingsdata niet representatief is van de gewenste taal en dat de grootte voor een training set niet groot genoeg is (Bisani and Ney, 2008) [7].

2.1.2.2 Prosodische analyse

Als laatste moet de prosodie worden berekend. Hierin wordt er bepaald waar de klemtonen op woorden worden gezet, welke woorden met extra nadruk worden uitgesproken, hoe de melodie en ritme van een zin verloopt, etc. Het berekenen van prosodie is zeer complex en bepaalt voor een groot deel ook de natuurlijkheid en verstaanbaarheid van een TTS systeem [8]. Bij deze laatste analyse wordt er voornamelijk gekeken naar drie aspecten van prosodie; prosodische structuur, accenten en melodie.

Tijdens het spreken worden sommige woorden gegroepeerd onder één adem, dit wordt ook wel prosodische frasering genoemd. Voor een TTS systeem is het belangrijk dat de zinnen in de juiste frase worden opgebroken. Het zorgt niet alleen dat een TTS systeem goed verstaanbaar is, maar het zorgt er ook voor dat de melodie en intonatie correct worden berekend. Intonatie is het verloop van toonhoogte in een uitgesproken zin. Er bestaan geen vaste regels wanneer een prosodische frase binnen een zin eindigt, maar vaak hangt dit samen met syntactische structuur. Om de zinnen in frases te splitsen, kan gebruik gemaakt worden van classificeerders of regel-gebaseerde algoritmes om te bepalen waar een frase eindigt zoals beschreven in Jurasfky en Martin (2009) [4] en Xu (2015) [9].

Bij deze analyse worden ook de accenten in de tekst berekend, zodat een zin niet monotoon wordt uitgesproken. Dit kan gedaan worden door klemtonen te leggen op lettergrepen in woorden. In zinnen geldt dat een accent ervoor zorgt dat een woord of zinsdeel in focus wordt gezet voor een luisteraar en zullen dan ook sneller opgemerkt worden door een luisteraar. De spreker wil aangeven dat het geaccentueerde gedeelte belangrijk is. Er kunnen vaste regels gebruikt worden om te bepalen welke woorden of lettergrepen een klemtoon krijgen. In het Nederlands wordt bijvoorbeeld 'nieuwe' informatie benadrukt bij het spreken doormiddel van accenten in zinnen. Men kan met dit soort regels een algoritme of programma schrijven, zodat de intonatie en accenten voor de rest van de tekst behandeld kunnen worden. In Quené en Kager (1989) [10] worden deze algoritmes voor prosodische analyse verder behandeld.

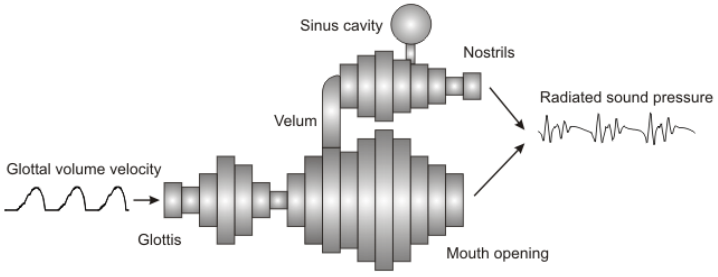
2.1.3 Spraakgenererende component

Nadat de tekst bewerkt en voorzien is van een prosodische beschrijving, wordt de spraak gesynthetiseerd in de spraakgenererende component. In deze component wordt geluid geproduceerd. Voor het genereren van spraak kunnen verschillende methodes gebruikt worden.

2.1.3.1 Articulatie synthese en Formant synthese

Articulatorische synthese is een type synthese die gebaseerd is op de werking van de menselijke glottis en de articulatoren. Synthetische spraak van articulatorische synthese wordt door middel van een computer gegenereerde spraakkanaal en stemplooien geproduceerd. De posities van de tong, lippen en kaak en de houding van de stembanden zijn parameters om verschillende klanken te kunnen maken. De luidheid en toonhoogte kunnen door middel van de trillende stembanden en verschil in luchtdruk van de luchtstroom beïnvloed worden. Dat wordt ook bij articulatorische synthese geproduceerd met een kunstmatige energiebron die de luchtstroom simuleert. Tijdens het produceren van spraak kunnen

de articulatoren manueel aangepast worden, zodat er meerdere klanken kunnen worden geproduceerd. Articulatorische synthese kan zeer natuurlijke en realistische spraak produceren (Levinson, 2012) [11]. Het wordt echter in praktijk minder gebruikt vanwege de complexiteit van het menselijke spraakorgaan. In figuur 2.2 is een voorbeeld te zien van een mogelijke opzet van een articulatie TTS systeem.



Figuur 2.2: Voorbeeld model van een articulatie synthesizer

Bij formant synthese wordt spraak geproduceerd met akoestische modellering met behulp van formanten. Formanten zijn boventonen of harmonischen, die door resonanties versterkt worden. Klanken hebben specifieke resonanties om zich te onderscheiden van andere klanken. Deze resonanties ontstaan bij de mens in de mond-keelholte, het holle gebied tussen de stembanden en mond. Het zorgt ervoor dat de korte luchtpulsen vanuit de stembanden waarin de grondtoon, ook wel F0 gerefereerd, en de harmonischen, tonen waarvan de frequentie de veelvoud is van de frequentie van de grondtoon, worden omgezet naar klanken. De resonanties versterken of verzwakken de boventonen. De vorm en de volume van de mond-keelholte, die kan veranderen door articulatoren, hebben invloed wat voor klank gemaakt kan worden. Er zijn 5 formanten die gerefereerd worden als F1 tot en met F5. De frequentiewaardes van F1 en F2 bepalen wat de klank kan worden (Rietveld, 2014) [3]. Bij deze synthese wordt er niet zozeer gekeken naar de werking van de spraakorgaan van een mens, maar wordt er meer gekeken naar de klanken en resonanties die met behulp van de spraakorgaan worden verkregen om spraak te creëren. Een probleem met deze synthese is dat het zeer lastig is om coarticulatie, een proces waarbij twee fonemen vloeiend naar elkaar overgaan, te modelleren vanwege de complexiteit.

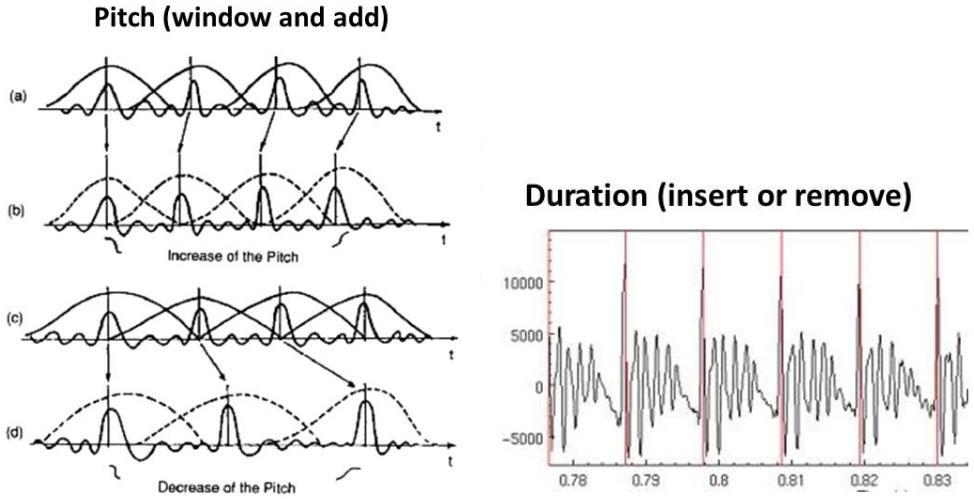
Articulatorische synthese en formant synthese maken dus gebruik van regels om spraak te produceren zonder vooropgenomen spraak in tegenstelling tot concatenatie synthese. De nadelen van deze regel gebaseerde synthese zijn dat het zeer ingewikkeld kan zijn om goede regels te maken en vooral dat de overgang tussen spraakklanken moeilijk te beregelen is wegens het vinden van de juiste parameters.

2.1.3.2 Concatenatie synthese

Bij concatenatie synthese wordt er gebruikt gemaakt van stukken opgenomen menselijke spraak. Een spraakdatabase van een concatenatie TTS systeem zal vele uren aan spraakopnames bevatten. Bij concatenatie synthese worden aan de hand van spraaksegmenten die het beste bij de symbolische klank beschrijving passen aan elkaar geplakt. Alleen moet de duur en toonhoogte altijd berekend worden. Op deze manier wordt spraak geproduceerd bij concatenatie synthese. Voor het produceren van spraak met concatenatie worden twee technieken vaak gebruikt [12].

Difoon concatenatie en PSOLA Om spraak te produceren met concatenatie synthese, kan er gebruik gemaakt worden van difonen. Difonen zijn overgangen van twee fonemen waarbij coarticulatie waarneembaar is. Een gemiddelde TTS systeem zou als er ongeveer

43 klanksegmenten bestaan, een combinatie van $43^2 = 1849$ difonen bevatten. Een professionele spreker spreekt dan tekst in waar de difonen uit worden geknipt. Deze worden dan opgeslagen in een database, zodat deze bij het maken van spraak aan elkaar geplakt kunnen worden. Voor elke difoon type is vaak maar één unit, oftewel spraakopname, aanwezig. De difonen worden gematcht met de fonetische beschrijving van de zinnen, vervolgens moeten de toonhoogte en duur verwerkt worden.



(a) Illustratie toonhoogte manipulatie (b) Illustratie duur manipulatie

Figuur 2.3: Illustratie van TD-PSOLA algoritme

Wanneer spraak wordt geproduceerd door difonen aan elkaar te plakken, moeten vervolgens de toonhoogte patroon en ritme gemaakt worden. Bij Difoon concatenatie moet de toonhoogte en duur altijd berekend worden. Dit kan met een algoritme genaamd TD-PSOLA gedaan worden. TD-PSOLA staat voor Time Domain Pitch Synchronous Overlap and Add. Het is een vaak gebruikt algoritme om de toonhoogte en duur van spraaksignalen te manipuleren. Ter illustratie van figuur 2.4, het idee van TD-PSOLA is om een gedeelte van de golfvorm waar een periode van de grondtoon F0 bevindt, genaamd een frame, te kopiëren en het te bewerken. Deze frames kunnen uitgestrekt of versmald worden zoals te zien in figuur 2.4a om een lagere of hogere toonhoogte te verkrijgen. Om de duur te manipuleren kunnen meerdere frames worden toegevoegd of verwijderd worden, zodat het fragment langer duurt. Meer exacte beschrijving van deze algoritme kan in Taylor (2009) [13] of in Jurafsky et Martin (2008) [4] gelezen worden.

Unit selection De meeste bedrijven maken gebruik van de methode Unit selection voor TTS systemen. Bij unit selection heb je minstens 10 uur aan spraak in de spraakdatabase zitten. Voor elke letter of grafeem is elke mogelijke klank, genaamd een unit, opgeslagen. Een voordeel van unit selection ten opzichte van difoon synthese is dat er bij unit selection weinig tot geen prosodische berekening gedaan hoeft te worden door de uitgebreide spraakdatabase. Het doel is om de beste reeks aan units op klank en prosodie te zoeken die correspondeert met de (prosodische) beschrijving van een zin.

Om te bepalen wat de beste reeks is, kan er gebruikt gemaakt worden van het Hunt and Black algoritme. Het Hunt and Black algoritme was gebruikt in één van de eerste unit selection systemen om Engelse spraak te synthetiseren (Taylor, 2009) [13]. Dit algoritme zoekt voor elke mogelijke reeks aan spraakeenheden naar de beste units. Zij speciferen de beste reeks als de reeks die de laagste waarde behaalt voor de Target cost en de Join cost functies. De Target cost functie berekent hoe goed de spraakeenheid matcht aan

de specificatie van een gevraagde fonetische en prosodische beschrijving. De Join cost functie toont aan of spraakeenheid S goed aansluit met de opvolgende spraakeenheid T. De join cost functie kan de waarde berekenen door de akoestische eigenschappen tussen de spraakeenheden te vergelijken.

$$C(U,S) = \sum_{t=1}^{T} T(u_t, s_t) + \sum_{t=1}^{T-1} J(u_t, u_{t+1})$$

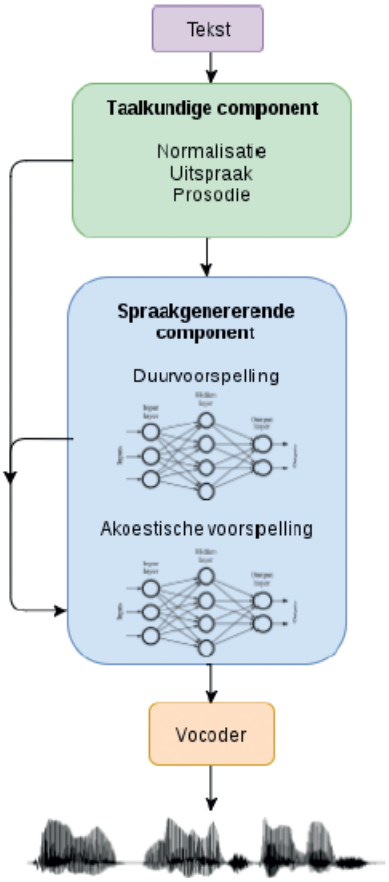
Figuur 2.4: Weergave wiskundige formule van totale kosten van een zin met Target en Join cost van het Hunt and Black algoritme

Deze kosten voor een hele zin wordt aan de hand van de functie in figuur 2.5 berekend. De waarde van de Target cost functie en de Join cost functie worden bij elkaar opgeteld voor alle mogelijke reeksen. De reeks die het laagste waarde behaalt, wordt gekozen als het beste reeks om spraak te produceren voor de gevraagde zin zoals beschreven in Hunt et Black (1996) [14] en Jurasfky et Martin (2008) [4]. Dit is één van de manieren hoe unit selection uitgevoerd kan worden. In Taylor (2009) [13] worden andere manieren van de methodes beschreven om de juiste units te kiezen. Een nadeel bij concatenative synthese dat de fragmenten niet goed op elkaar aansluiten, wat leidt dat de melodie niet goed zal verlopen.

2.1.3.3 Deep Neural Network (DNN)

Een opkomende techniek voor het produceren van synthetische spraak is met behulp van deep neural networks. Tegenwoordig zijn computers zelf in staat om automatisch te leren met machine learning. Zo zijn spamfilters getraind om spammail te herkennen en automatisch te verplaatsen naar de spam folder. Machine learning zou ook toegepast kunnen worden op TTS systemen. Volgens Klabbers (2017) [12] zouden er twee DNN modellen getraind moeten worden om spraak te genereren, een DNN model om de duur van klanken te berekenen en een DNN model om fonemen om te zetten naar de akoestische eigenschappen. De vocoder (Voice Encoder) krijgt deze informatie binnen en zet het uiteindelijk om in spraak. Een nadeel bij DNN TTS systemen is dat er relatief veel tijd en training nodig is om spraak natuurlijk en duidelijk te produceren.

Een voorbeeld van deze techniek is terug te zien bij de bedrijven Apple en Google. Apple en Google maken al gebruik van deep neural network (DNN) om spraaksynthese nog natuurlijker te laten klinken. Een concreet voorbeeld hiervan is WaveNet van DeepMind. Het TTS systeem is met behulp van convolutionele neurale netwerken gebouwd, afgekort (CNN). CNN is een type neurale netwerk waarbij niet alle neuronen met elkaar verbonden zijn, wat normaal bij een neurale netwerk vaak het geval is, en waar de input door verschillende lagen van het netwerk gefilterd wordt totdat er een gewenste



Figuur 2.5: Basis opzet van een DNN TTS Systeem, *bron: Klabbers, E. Dixit, 2017*

feature eruit gehaald kan worden. Zo bestaat een CNN uit een input layer, een output layer en verscheidene hidden layers. Volgens van Den Oord et al. (2016) [15] is gebleken dat WaveNet uitstekend presteerde om natuurlijk te klinken uit een subjectieve evaluatie. Meer exacte informatie over de implementatie en werking van WaveNet kan gelezen worden in van Den Oord et al. (2016) [15].

2.2 Evaluatie van spraaksynthese

Om verbetering van TTS systemen daadwerkelijk vast te stellen, zouden er evaluaties moeten worden uitgevoerd om te bepalen hoe een TTS systeem presteert. Dit kan zowel objectief of subjectief uitgevoerd worden.

Voor bedrijven kost een subjectieve evaluatie echter veel tijd vanwege het opzetten van het experiment, het werven van proefpersonen en het afhandelen van vergoedingen aan proefpersonen. Het maakt dit soort onderzoek ingewikkeld om het uit te voeren. Daarnaast is er de mogelijkheid dat TTS systemen objectief te beoordelen. Synthetische stemmen kunnen door methodes zoals Perceptual Evaluation of Speech Quality (PESQ) beoordeeld worden, wat bijvoorbeeld oom gebruikt wordt om de spraakkwaliteit van telefonie te beoordelen (Rix et al., 2002) [16].

Dit betekent niet dat de resultaten van objectieve evaluaties altijd in overeenstemming zijn met resultaten van subjectieve evaluaties. Er wordt voorlopig aangenomen dat bedrijven meer gebruik maken van objectieve evaluaties vanwege de tijdsduratie en relatief lage kosten. Voor bedrijven zoals ReadSpeaker zou het interessant zijn om resultaten uit subjectieve evaluaties te verkrijgen, want de resultaten zouden voor dit soort bedrijven kunnen leiden tot andere inzichten en verbeterpunten die niet met objectief onderzoek gevonden worden.

2.2.1 Subjectieve evaluatie

Een subjectieve evaluatie kan op verschillende manieren worden uitgevoerd. Proefpersonen kunnen een TTS systeem beoordelen op verstaanbaarheid met rijmtesten, **Diagnostic Rhyme Test (DRT)**, of luisteropdrachten waarbij de proefpersoon de zin moet overschrijven wat hij of zij gehoord heeft. Er wordt dan gekeken hoe goed proefpersonen het TTS systeem verstaan. Er kan echter ook beoordeeld worden aan de hand van opinie. Hiervoor kunnen testen worden gebruikt waar proefpersonen een score geven of een preferentie van een spraaksegment doorgeven (Cryer et Home, 2010) [17]. Hier worden twee soorten testen besproken.

Mean opnion score (MOS) Bij deze test wordt algemeen de prestatie van een TTS systeem beoordeeld. Een proefpersoon krijgt verschillende gesynthetiseerde stukken spraak van 3 tot 5 zinnen te horen waar er een score wordt toegekend aan gevraagde features tussen 1 tot 5. Er kan dan gevraagd worden om een score te geven op de attributen verstaanbaarheid, natuurlijkheid, melodie, prettigheid van stem, etc. De TTS systemen kunnen met elkaar vergeleken worden met dezelfde testzinnen.

AB test Bij deze test wordt er van een proefpersoon gevraagd om een voorkeur te kiezen. In een AB test wordt er voor de proefpersoon twee synthetische spraaksegmenten, fragment A en B, afgespeeld van twee verschillende TTS systemen. De proefpersoon kiest het fragment die hij of zij het beste vindt. Hiervoor worden er minstens 50 zinnen voor gebruikt. De voorkeurzinnen van de TTS systemen worden dan met elkaar vergeleken. Om bias te voorkomen, worden de synthetische spraakfragmenten op willekeurige volgorde gegeven (Jurasfky en Martin, 2008) [4].

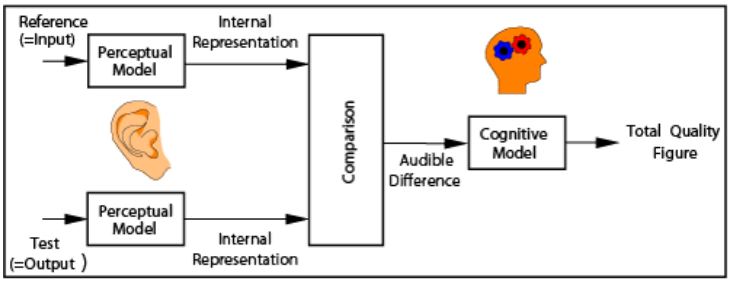
Een concrete voorbeeld van waar subjectieve evaluaties worden uitgevoerd, is The Blizzard Challenge. The Blizzard Challenge is een jaarlijkse opdracht waarbij teams van onderzoekers van universiteiten of instituten uit verschillende landen een nieuwe stem ontwikkelen op basis van een vrijgegegeven spraakdatabase van de organisatie achter The Blizzard Challenge. Hierin zijn wetenschappers vrij om te kiezen met behulp van

welke technieken een synthetische stem wordt gemaakt. Om de stemmen te beoordelen, maakt de organisatie van The Blizzard Challenge gebruik van luistertesten. In 2017 bestonden deze luistertesten uit 7 secties waarvan er op elke sectie 16 tot 17 samples te beluisteren waren. De proefpersonen moesten een score tussen 1 en 5 geven op elk van de attributen algemene presentatie, aangenaamheid, spreekpauzes, intonatie, emotie, klemtoon en luisterinspanning. Deze groep proefpersonen bestond uit betaalde studenten van Edinburgh University tussen de 18 en 25 jaar die Engels als moedertaal beheersen, spraakdeskundigen en vrijwilligers die via de deelnemers van The Blizzard Challenge zijn gerekruteerd zoals beschreven in King et al. (2017) [18].

2.2.2 Objectieve evaluatie

Een andere manier om de kwaliteit van spraaksynthese te meten is door objectief methodes te gaan gebruiken. Momenteel wordt er veel onderzoek verricht om een objectieve methode te vinden waarbij de resultaten van subjectieve evaluaties één-op-één kunnen staan met objectieve evaluaties. Een objectieve methode die gebruikt wordt om de kwaliteit van synthetische spraak te meten is de methode PESQ.

PESQ staat voor Perceptual Evaluation of Speech Quality. Het is een automatische meetmethode om spraak van een telefoonsysteem te beoordelen gebaseerd op de subjectieve methode MOS. In figuur 2.4 is de opzet van het methode te zien. Kort samengevat, PESQ krijgt twee geluidsfragmenten binnen in zijn perceptuele modellen, een referentie fragment en een testfragment. De testfragment wordt aan de aan de hand van een referentiefragment vergeleken op juistheid van de zin, verstaanbaarheid en toonhoogte. De verschillen worden dan doorgegeven naar het cognitieve model en uiteindelijk wordt er een score bepaald. Voor een meer technische achtergrond zie Beerends et al. (2002) [16].



Figuur 2.6: Basis model van PESQ methode

In Cernak en Rusko (2005) [19] werd gebruikt gemaakt van PESQ om spraak te beoordelen. Tijdens hun onderzoek verrichtte ze een subjectieve en objectieve evaluatie om vervolgens een correlatie te berekenen tussen de behaalde resultaten. Drie verschillende TTS systemen moesten 10 lijsten van 10 Slowaakse woorden uitspreken. De proefpersonen moesten de gesynthetiseerde spraaksegmenten beoordelen met de methode MOS. De PESQ methode kreeg als referentie de menselijke spraaksegmenten om het te vergelijken met de synthetische spraaksegmenten. Uit het onderzoek kwam een hoge correlatie tussen PESQ resultaten en subjectieve resultaten. Er werd echter nog vermeld dat deze methode niet werkt voor een te kleine test set en dat de methode niet direct gebruikt kan worden voor evaluaties zonder eerst lineaire transformaties te doen, aangezien de PESQ scores ten opzichte van de MOS scores gehalveerd waren. In Rix et al.(2001) [20] werd er ook vastgesteld dat er een hoge correlatie is tussen PESQ en subjectieve resultaten en dat de PESQ methode bruikbaar is voor het meten van synthetische spraak kwaliteit.

3 Opbouw van onderzoek

In dit hoofdstuk wordt de onderzoeksvraag met bijbehorende deelvragen gesteld. Tevens wordt de opzet van het experiment beschreven en worden de gemaakte keuzes onderbouwd met argumentatie.

3.1 Onderzoeksachtergrond

Voor dit onderzoek werd er gekozen om een subjectieve evaluatie uit te voeren om te achterhalen wat de gemiddelde beoordeling is van een gebruiker.

Bij dit soort evaluaties zijn factoren betrokken waarop er gelet moet worden. In Wester et al. (2015) [21] wordt er checklist beschreven waar een subjectieve evaluatie zich moet houden. Er moet gelet worden: 1) welke vragen je gaat stellen, aangezien de formulering van een vraag voor elke proefpersoon anders geïnterpreteerd kan worden, 2) of het aantal proefpersonen voldoende is om statistische uitspraken te kunnen maken, 3) welke subjectieve test gebruikt gaat worden, 4) of deze evaluatie online of fysiek wordt afgenomen, etc.

Uit hun onderzoek is ook gebleken dat een aantal van minstens 30 luisteraars voor een MOS test meer betrouwbaarheid. Voor het onderzoek van deze scriptie werd zoveel mogelijk aan de werkwijze van de jaarlijkse Blizzard Challenge gevolgd en aan de checklist die beschreven staat in Wester et al. (2015) voldaan.

Er werd gekozen voor een MOS test die op één laptop met speakers zonder koptelefoon afgenomen wordt. De MOS test werd eerder in de sectie Theoretische kader behandeld. De luisteromgeving wordt zoveel mogelijk onder controle gehouden, zodat . Een exacte opzetbeschrijving van het experiment met deze methode wordt later beschreven in de sectie Methode.

3.2 Experimentele opzet

3.2.1 Proefpersonen

Het experiment beperkt zich tot de doelgroep universitaire en HBO studenten die Nederlands als moedertaal beschikken. Deze groep is in staat om kritische meningen en feedback te geven over de geproduceerde spraak van verschillende TTS systemen.

3.2.2 Stemmen

Voor dit onderzoek werden in totaal 8 synthetische stemmen van verschillende bedrijven en een menselijke mannenstem als controlestem gebruikt. De meeste TTS systemen voor dit onderzoek maken gebruik van de techniek Unit selection die eerder werd uitgelegd in de sectie Theoretische kader. De controlestem is een 21-jarige mannelijke universitaire student. De stemmen 2 tot en met 8 zijn verkregen uit de tool ReadSpeaker Speechmaker. Readspeaker Speechmaker is een tool waarbij je audio bestanden kan maken van teksten die omgezet worden naar spraak. De tekst kan door verschillende TTS systemen worden ingesproken en de gemaakte audio bestanden kunnen vervolgens gedownload worden. Verder is het met Readspeaker Speechmaker mogelijk om de snelheid en toonhoogte van spraak aan te passen en men kan zelf uitspraken van woorden, acroniemen etc. vast leggen in een gepersonaliseerde woordenboek. De ResponsiveVoice stem is verkregen van ResponsiveVoice developer API. ResponsiveVoice is een onderdeel van het bedrijf Learnbrite.

Nr.	TTS systeem	Nr.	TTS systeem
1	ResponsiveVoice Google stem	5	Acapela Femke
2	ReadSpeaker Guus	6	Readspeaker Ilse
3	Fluency Arno	7	Nuance Xander
4	Nuance Claire	8	Ivona Ruben

Tabel 3.1: De namen en bijbehorende bedrijven van TTS systemen die worden onderzocht.

Arno De stem Arno bezit een spraakdatabase van bijna 3 uur en ongeveer 90.000 spraakklanken. Het is afkomstig van Fluency en maakt gebruik van unit selection [22].

Femke Deze vrouwelijke stem was als enige beschikbaar van het bedrijf Acapela. Voor diversiteit werd het TTS systeem meegenomen in het onderzoek. Maakt gebruik van unit selection [23].

Ruben De mannenstem Ruben kan je bijvoorbeeld terug horen op Amazon Polly. Vanwege dezelfde reden als Femke, werd deze stem meegenomen in het onderzoek. Maakt gebruik van unit selection [24].

Ilse De vrouwenstem Ilse was als enige stem van Readspeaker beschikbaar in het begin van het onderzoek. De stem is gebaseerd op unit selection.

Guus De mannenstem Guus is een vrij nieuwe stem van ReadSpeaker die nog uitgebracht moet worden. De stem was tijdens het onderzoek nog in ontwikkeling, maar was al in goede staat om geëvalueerd te worden. Voor het bedrijf en het onderzoek was het interessant om te bekijken hoe de stem Guus presteert. De stem is gebaseerd op unit selection.

Claire De vrouwenstem van Nuance klonk als een rustige en vloeiende stem. Maakt gebruik van unit selection [25].

Xander De mannenstem van Nuance klonk relatief vloeiend en helder. Maakt gebruik van unit selection [25].

Google stem De meeste mensen kennen de service Google Translate wel. Hiervan is er een kunstmatige stem aanwezig voor het Nederlands. Het bedrijf biedt geen informatie over de implementatie van de stem.

3.2.3 Teksten

Voor het experiment experiment werden 5 teksten gekozen. Het zijn teksten waarmee studenten dagelijks in aanraking komen. Denk hierbij aan nieuwsartikelen, literatuur of handleidingen. De volgende teksten zijn voor het onderzoek gebruikt.

Tekst	Soort tekst	Aantal woorden	Gemiddelde spreekduratie
1	Nu.nl Nieuwsartikel	83	30s
2	Interview	81	30s
3	Onderzoeksgids Geschiedenis	97	34s
4	Handout Wiskunde	109	45s
5	Nu.nl Nieuwsartikel	97	38s

Tabel 3.2: Specificatie van testteksten. Volledige teksten zijn in Bijlage A te vinden.

3.2.4 Gestelde vragen

Voor dit experiment werden de volgende vragen gesteld. De vragen over verstaanbaarheid, natuurlijkheid, articulatie en melodie zijn afgeleid van de type vragen van de jaarlijkse Blizzard Challenge.

Van welke studie komt de proefpersoon? Studenten die van een talenstudie komen, hebben mogelijk meer kennis over spraak dan studenten van andere studies. Tevens geldt dit ook voor studies zoals informatica waar men hoogstwaarschijnlijk een kunstmatige stem heeft gehoord. Deze vraag dient meer als achtergrondkennis van de luisteraar. Dit is een open vraag.

Heeft een proefpersoon eerder een kunstmatige stem gehoord? Er kan een verschil in resultaten optreden tussen een persoon die eerder een TTS systeem gehoord heeft en een proefpersoon die een TTS systeem voor het eerst hoort. Dit is een multiple choice vraag.

Vond je deze stem passen bij de tekst? Elke TTS systeem is anders. Sommige systemen passen beter bij nieuwsartikelen en andere bij korte mededelingen. Daarom bekijken we ook of bepaalde stemmen bij een bepaalde soort tekst beter passen vanwege karakter of uitspraak. Dit is een ja-nee-vraag.

Bij deze vraag was er waarom-vraag aanwezig, zodat een proefpersoon zijn keuze kon onderbouwen.

Hoe verstaanbaar vind je deze stem? Deze vraag is noodzakelijk om erachter te komen of een stem überhaupt verstaanbaar is. In praktijk zou een stem niet worden gebruikt als die niet wordt verstaan. Dit is een lineaire schaalvraag met de waarde van 1 tot 5.

Hoe prettig vind je de melodie van deze stem? De intonatie van een stem speelt een belangrijke rol of een TTS systeem voor de luisteraar verdraaglijk is om een tekst te beluisteren. De stem moet niet te monotoon zijn, want er is dan kans dat je op deze manier de concentratie en focus van een luisteraar sneller verliest. Dit is een lineaire schaalvraag met de waarde van 1 tot 5.

Hoe duidelijk vond je de articulatie van deze stem? De proefpersoon beoordeelt hoe duidelijk de afzonderlijke spraakklanken te horen zijn. Een slechte articulatie kan leiden naar slechte verstaanbaarheid. Door deze vraag te stellen, kan er herleid worden of een lage score aan verstaanbaarheid veroorzaakt wordt door de articulatie of dat het door een andere factor wordt veroorzaakt. Dit is een lineaire schaalvraag met de waarde van 1 tot 5.

Hoe natuurlijk vind je deze stem klinken? Deze vraag wordt gesteld in geval dat deze factor invloed heeft op de resultaten van andere vragen. Als de natuurlijkheid van een stem lager is, kan het zijn dat de stem minder goed verstaanbaar is of lager wordt beoordeeld op verstaanbaarheid of melodie. Dit is een lineaire schaalvraag met de waarde van 1 tot 5.

Hoe waarschijnlijk lijkt het je dat je deze stem zou gebruiken om teksten voor te lezen? Deze vraag schetst of proefpersonen het aannemelijk vinden dat zij zelf deze stemmen zullen gebruiken voor persoonlijke doeleindes. Dit is een multiple choice vraag. Bij deze vraag was er waarom-vraag aanwezig, zodat een proefpersoon zijn keuze kon onderbouwen.

Denk je dat deze stem je zou helpen sneller informatie uit teksten op te nemen dan wanneer je de tekst zelf leest? Deze vraag wordt gesteld of spraaksynthese een mogelijke positieve werking kan geven op een gebruiker. Dit is een multiple choice vraag.

Welke fragment vond je het best? Deze vraag dient ervoor of een proefpersoon consistent is geweest met zijn antwoorden aan welke deze persoon de beste score heeft gegeven. De proefpersoon kiest één fragment van de 5 teksten. Dit is een multiple choice vraag.

3.2.5 Opzet

Er werden acht synthetische stemmen en één menselijke mannenstem willekeurig verdeeld in twee groepen van vijf, wat leidt dat één stem in beide groepen voorkomt. De reden dat er twee groepen van vijf werden gemaakt is, zodat een luisterexperiment uitvoerbaar is voor studenten met relatief weinig tijd. Vaak is het geval dat een student relatief weinig tijd heeft wegens werk, verplichtingen, sociale leven etc. De kans dat studenten een luisterexperiment weigeren zou groter zijn wanneer het experiment relatief gezien erg lang duurt. Aangezien het om een test gaat waar fragmenten beluisterd moeten worden en dat er relatief veel vragen worden gesteld, was het gunstiger om twee groepen te maken met 5 stemmen. De tijd om een luisterexperiment te ondergaan werd gemeten op advies van Wester et al. (2015). Gemiddeld duurt het luisteren naar fragmenten en het invullen

van vragen 20 minuten voor de proefpersonen wanneer een luisterexperiment 5 audio bestanden bevat.

Groep 1 bevat 5 synthetische stemmen en groep 2 bevat één menselijke stem en vier synthetische stemmen. De verdeling is zo gemaakt dat er relatief gezien evenveel mannen- als vrouwenstemmen aanwezig zijn in één groep.

Nr.	TTS systeem Groep 1	TTS systeem Groep 2
1	RSPV stem	Guus
2	Arno	Claire
3	Femke	Ilse
4	Ilse	Xander
5	Ruben	Controlestem

Tabel 3.3: Enquête indeling van TTS-systemen. RSPV is de afkorting van ResponsiveVoice. In beide groepen komt de TTS systeem Ilse voor om een gelijke verdeling te verkrijgen en Groep 2 bevat een menselijke controlestem. Dit wordt uitgelegd in hoofdstuk 5 Discussie

Het experiment is opgesteld met behulp van Google Forms. Aangezien er gebruikt gemaakt werd van audio-bestanden, was er geen mogelijkheid in Google Forms om audio af te spelen via een audiospeler. Er was een optie geweest om de .wav-bestanden te converteren naar .mp4-bestanden om het dan via de service Youtube als een video in de enquête direct af te kunnen spelen, maar dat zou geen garantie geven dat de audio bestanden dezelfde kwaliteit zouden behouden na het converteren. Er werd daarom gekozen om een aparte weblink te maken die naar de audio bestanden leidde, zodat deze bestanden zo min mogelijk worden gemanipuleerd. Deze enquêtes werden via een laptop voor elke proefpersoon op constante volume in een stille luisteromgeving afgenomen.

Een enquêtelijst bestaat uit 3 pagina's waar instructies staan en 7 pagina's met vragen. De eerste vragenpagina bevat vragen om achtergrondkennis van de luisteraar te verkrijgen. De laatste vragenpagina bevat één vraag die te maken heeft met de 5 geluidsfragmenten. De resterende vragenpagina's bevatte geluidsfragmenten van de testteksten, ingesproken door een andere stem, en een lijst met vragen die beschreven staat in sectie 3.4.3. De lijst bestond uit 9 vragen; 5 lineaire schaalvragen, 2 multiple choice vragen en 2 open vragen.

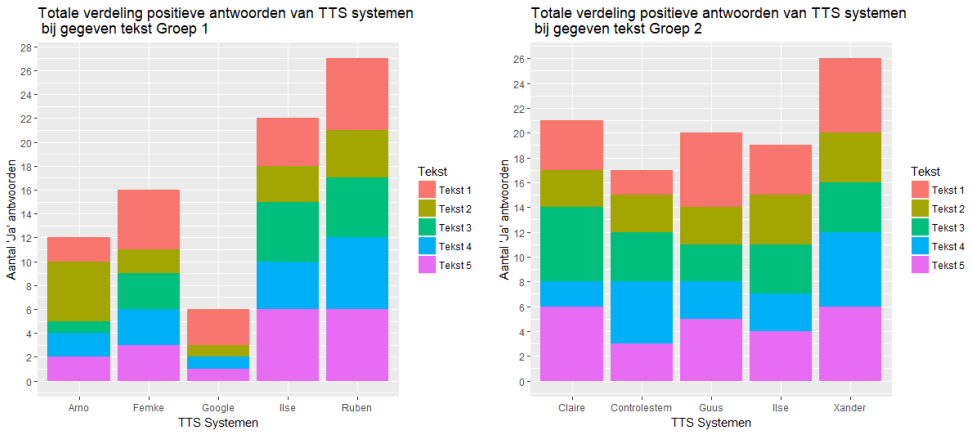
Er zijn in totaal 10 verschillende enquêteversies. Elk van de twee groep heeft 5 verschillende enquête versies. De reden om voor beide groepen 5 verschillende enquêteversies te maken is dat elke stem wordt beoordeeld aan de hand van de 5 verschillende teksten. Het kan zijn dat stem 1 erg goed wordt ervaren in het voorlezen van nieuwsartikelen, terwijl stem 2 relatief slecht wordt ervaren in het voorlezen van literaire artikelen. Het zou dus zeer eenzijdig en onbetrouwbaar zijn om een stem te beoordelen waar het relatief erg goed of slecht op presteert.

De opzet van de teksten staan in Bijlage B. De volgorde van de voorgelezen teksten is voor elke enquête hetzelfde, maar deze wordt afhankelijk van testversie voorgelezen door een andere stem. Verder konden proefpersonen een audio-fragment eenmalig beluisteren.

4 Resultaten

Er waren in totaal 60 Nederlandse proefpersonen van Universiteit Utrecht, Hogeschool Utrecht, Universiteit Rotterdam en Technisch Universiteit Eindhoven tussen de 18 tot 26 jaar, betreffende 38 mannen en 22 vrouwen, die hebben meegedaan aan dit onderzoek. 30 personen voor groep 1 en 30 personen in groep 2. Deze proefpersonen waren verkregen door middel van reclame, vrienden en familie. De rekenkundige gemiddeldes en standaardafwijkingen zijn verkregen door de software Rstudio. De principale componenten matrix is met behulp van het programma SPSS Statistics verkregen. De variantieanalyse en chi-kwadraattoets zijn doormiddel van de software Microsoft Excel uitgevoerd.

Tijdens de enquête werd als eerste de vraag gesteld welke opleiding de proefpersoon volgt. In Bijlage D is een overzicht van opleidingen te zien. Hierin is te zien dat de proefpersonen afkomstig zijn van verschillende opleidingen. Deze gegevens worden verder niet gebruikt. De tweede vraag die werd gesteld of een proefpersoon eerder een kunstmatige stem heeft gehoord. Hierop hadden 53 mensen 'Ja' geantwoord, 5 met 'Misschien' en 2 met 'Nee'. De meeste luisteraars die in het proef hebben meegedaan hebben dus eerder een kunstmatige stem gehoord.



(a) Resultaten Groep 1

(b) Resultaten Groep 2

Figuur 4.1: Resultaten van de vraag: "Vond je deze stem passen bij de tekst?" van proefpersonen uit groep 1 en 2. De kleuren geven aan uit welke teksten er een positieve antwoord werd gegeven voor verschillende TTS systemen.

In figuur 4.1 is er een verdeling te zien van hoeveel positieve antwoorden op de vraag of een TTS systeem bij een tekst past. Deze resultaten komen van de eerste vraag die in Bijlage C in de figuren C.2, C.3a en C.3b te zien zijn. De verschillende kleuren geven aan om welke testtekst het gaat. In figuur 4.1a is er te zien dat de ResponsiveVoice stem weinig ja-antwoorden kreeg zowel in zijn eigen groep als van de totale groep. De controlestem heeft binnen groep 2 ook het laagste aantal positieve antwoorden, ondanks dat de controlestem een menselijke stem is. De mannelijke kunstmatige stemmen Ruben en Xander hebben het hoogste aantal positieve antwoorden verkregen van alle typen teksten.

Om te achterhalen of de verschillen in het totaal aantal ja-antwoorden tussen de TTS systemen niet op toeval berust, wordt de methode chi-kwadraattoets ($\tilde{\chi}^2$) uitgevoerd. De

chi-kwadraattoets is een statistische toets om verschil aan te tonen tussen verschillende groepen. Het kan ook gebruikt worden om samenhang tussen twee variabelen aan te duiden. In dit geval wordt er slechts gekeken of de gevonden verschillen tussen de totaal aantal positieve antwoorden van TTS systemen significant zijn. Dit kan gedaan worden met de volgende formule:

$$X^2 = \sum (W - V)^2 / V$$

Figuur 4.2: Weergave van chi-kwadraat formule.

Hierbij is W de waargenomen waarde en V de verwachte waarde. Alle mogelijkheden van $(W - V)^2$ worden bij elkaar opgeteld en daarna gedeeld door de verwachte waarde. Vervolgens komt er een waarde χ^2 wat ook wel de toetsingsgrootheid genoemd wordt. Wanneer de χ^2-waarde gelijk of groter is dan de kritieke waarde, kan er gesteld worden dat de gevonden verschillen significant zijn. De kritieke waarde moet worden opgezocht in een tabel en hangt af van de α en het aantal vrijheidsgraden (ook wel df afgekort van het Engelse woord *degrees of freedom*). De gebruikelijke α waarde is 0.05 en de vrijheidsgraden wordt verkregen door het aantal groepen met 1 af te trekken (NIST/SEMATECH, 2002) [26]. Om de berekening van chi-kwadraat te vereenvoudigen, wordt er voor dit onderzoek aangenomen dat beide groepen luisteraars gelijk oordelen.

TTS systeem	Aantal Ja-antwoorden	TTS systeem	Aantal Ja-antwoorden
RSPV stem	6	Guus	20
Arno	12	Claire	21
Femke	16	Ilse ver. G	19
Ilse ver. B	19	Xander	26
Ruben	27	Controle	17

Tabel 4.1: Verdeling van positieve antwoorden over alle systemen. Ilse ver.B verwijst naar TTS systeem Ilse in groep 1 en Ilse ver.G verwijst naar TTS systeem Ilse in groep 2.

In tabel 4.1 is een weergave te zien van het aantal positieve antwoorden. De verwachte waarde V is het totaal aantal positieve antwoorden gedeeld door het aantal systemen wat uitkomt op 18.6. Met behulp van de X^2-formule wordt de volgende verkregen:

$$X^2 = \frac{(-12.6)^2 + (-6.6)^2 + (-2.6)^2 + (3.4)^2 + (8,4)^2 + (1,4)^2 + (2,4)^2 + (0.4)^2 + (7.4)^2 + (-1.6)^2}{18.6}$$

Uit de berekening volgt dat $\chi^2 = 17{,}19784946$. Om vast te stellen dat de verschillen significant zijn, moet de χ^2-waarde de kritieke waarde bij $\alpha = 0.05$ en df = 9 overtreffen. Uit figuur E.1 in Bijlage E is te zien dat de waarde 16.919 is. De waarde van χ^2 is groter dan de kritieke waarde, dus kan er vastgesteld worden dat de gevonden verschillen tussen de TTS systemen significant zijn.

Deze uitkomst vertelt echter niet welke TTS systemen onderling verschillen. Om dit uit te vinden wordt de chi-kwadraattoets uitgevoerd in paren van twee TTS systemen. De nieuwe verwachte waarde V wordt de rekenkundige gemiddelde van de desbetreffende gepaarde TTS systemen. De kritieke waarde voor paren TTS systemen met $\alpha = 0.05$ en df = 1 is 3.841 volgens figuur E.1.

Gepaarde TTS systemen	χ^2-waarde	Gepaarde TTS systemen	χ^2-waarde
RSPV stem-Femke	4.545	RSPV stem- Ilse ver.G	6.760
RSPV stem-Ilse ver.B	9.143	RSPV stem-Xander	12.500
RSPV stem-Ruben	13.364	RSPV stem-Controle	5.261
RSPV stem-Guus	7.538	Arno-Ruben	5.769
RSPV stem-Claire	8.333	Arno-Xander	5.158

Tabel 4.2: χ^2-waardes van gepaarde TTS systemen die de kritieke waarde bij $\alpha = 0.05$ en df = 1 overtreffen.

In tabel 4.2 wordt er alleen aangetoond welke paren TTS systemen de kritieke waarde overtreffen. Er kan gesteld worden dat een tweetal van TTS systemen significant van elkaar verschillen. In dit geval verschilt het TTS systeem van ResponsiveVoice met de rest van de TTS systemen op het systeem Arno na. Het systeem Arno verschilt met Ruben en Xander. Een meerendeel van de paren bevat het systeem van ResponsiveVoice. Dit duidt erop dat dit systeem het meest afwijkt ten opzichte van de rest wat ook terug te zien is in figuur 4.1, dus er kan gesteld worden dat het systeem van ResponsiveVoice het laagst heeft gepresteerd op deze vraag. Voor het TTS systeem Arno kan dit een lichte aanduiding zijn.

TTS systeem	Articulatie	Melodie	Natuurlijkheid	Verstaanbaarheid
RSPV stem	3.10	1.87	1.60	3.30
Arno	2.63	2.07	1.67	2.77
Femke	2.80	2.87	2.10	3.07
Ilse	3.63	3.27	2.80	3.90
Ruben	3.50	3.53	3.17	3.93

Tabel 4.3: Groep 1 gemiddelde μ MOS score voor elke feature

TTS systeem	Articulatie	Melodie	Natuurlijkheid	Verstaanbaarheid
RSPV stem	0.88	0.94	0.72	0.95
Arno	0.96	1.20	0.92	1.04
Femke	1.19	1.04	1.12	1.04
Ilse	0.89	0.91	1.27	0.66
Ruben	1.17	1.07	1.21	0.98

Tabel 4.4: Groep 1 standaarddeviatie of σ-waardes voor elke feature

TTS systeem	Articulatie	Melodie	Natuurlijkheid	Verstaanbaarheid
Guus	3.57	2.23	2.03	3.77
Claire	3.30	2.50	2.03	3.70
Ilse	3.50	2.83	2.37	3.97
Xander	3.33	3.03	2.70	3.80
Controle	3.37	3.33	4.13	4.03

Tabel 4.5: Groep 2 gemiddelde μ MOS score voor elke feature

TTS systeem	Articulatie	Melodie	Natuurlijkheid	Verstaanbaarheid
Guus	1.22	1.07	1.13	1.28
Claire	1.20	1.17	1.10	0.99
Ilse	1.04	1.15	1.33	0.85
Xander	1.06	1.00	1.09	0.92
Controle	1.16	1.21	0.97	0.96

Tabel 4.6: Groep 2 standaarddeviatie of σ-waardes voor elke feature

Tabel 4.3 geeft de rekenkundige gemiddeldes van TTS systemen weer van groep 1. Tabel 4.4 geeft de standaarddeviaties die bij groep 1 horen weer. Tabel 4.5 en 4.6 geven dit ook weer voor groep 2. De standaarddeviaties geven de spreiding aan scores aan vanaf de rekenkundige gemiddelde. 68% van de MOS scores zullen per onderdeel tussen de waarde μ-σ en μ+σ liggen.

Uit tabel 4.3 kan opgemaakt worden dat de TTS systeem Ruben een relatief hoge score bevat op de onderdelen articulatie, melodie en verstaanbaarheid. Verder is er in tabel 4.4 en 4.6 de standaarddeviaties te zien van desbetreffende systemen. De σ-waardes liggen tussen de 0.90 en 1.20, wat relatief groot is.

Ook hier wordt er nagegaan of de verschillen tussen de rekenkundige gemiddeldes significant zijn. Hiervoor wordt variantieanalyse, oftewel One-way ANOVA toegepast. De One-way ANOVA is een statistische toets dat gebruikt wordt om het verschil tussen tenminste 3 groepen als significant aan te tonen op basis van één factor. Er wordt bij deze methode gekeken naar de rekenkundige gemiddeldes van de groepen. Net zoals bij de chi-kwadraattoets wordt er bij de One-way ANOVA een toetsingsgrootheid F berekend. Wanneer de F-waarde de kritische gebied van F-toets overschrijdt, kan er gesproken worden om een significant verschil tussen de groepen (Bluman, 2009) [27].

Deze methode wordt door middel van een ingebouwde functie genaamd Unifactoriële variantie-analyse van Microsoft Excel uitgevoerd. Hier wordt ook voorlopig aangenomen dat beide groepen luisteraars gelijk oordelen om de berekening te vergemakkelijken.

Unifactoriële variantie-analyse

SAMENVATTING ARTICULATIE

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	30	93	3,1	0,782759
Arno	30	79	2,633333333	0,929885
Femke	30	84	2,8	1,406897
Ilse ver. Blauw	30	109	3,633333333	0,791954
Ruben	30	105	3,5	1,362069
Guus	30	99	3,3	1,458621
Claire	30	75	2,5	1,362069
Ilse ver. Groen	30	105	3,5	1,086207
Xander	30	100	3,333333333	1,126437
Controle	30	101	3,366666667	1,343678

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	41,8	9	4,644444444	3,986451	8,14E-05	1,912235867
Binnen groepen	337,8666667	290	1,165057471			
Totaal	379,6666667	299				

Figuur 4.3: Resultaten van One-way ANOVA in excel voor de lineaire schaalvraag "Hoe duidelijk vond je de articulatie van deze stem?". De F-waarde is 4.644 en overschrijdt het kritische gebied van F-toets.

Unifactoriële variantie-analyse

SAMENVATTING MELODIE

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	30	56	1,866666667	0,878161
Arno	30	62	2,066666667	1,443678
Femke	30	86	2,866666667	1,085057
Ilse ver. Blauw	30	98	3,266666667	0,822989
Ruben	30	106	3,533333333	1,154023
Guus	30	68	2,266666667	1,236782
Claire	30	75	2,5	1,362069
Ilse ver. Groen	30	85	2,833333333	1,316092
Xander	30	91	3,033333333	0,998851
Controle	30	100	3,333333333	1,471264

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	85,93666667	9	9,548518519	8,113303	9,66E-11	1,912235867
Binnen groepen	341,3	290	1,176896552			
Totaal	427,2366667	299				

Figuur 4.4: Resultaten van One-way ANOVA in excel voor de lineaire schaalvraag "Hoe prettig vind je de articulatie van deze stem?". De F-waarde is 9.549 en overschrijdt het kritische gebied van F-toets.

Unifactoriële variantie-analyse

SAMENVATTING NATUURLIJKHEID

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	30	48	1,6	0,524137931
Arno	30	50	1,666666667	0,850574713
Femke	30	63	2,1	1,265517241
Ilse ver. Blauw	30	84	2,8	1,613793103
Ruben	30	95	3,166666667	1,454022989
Guus	30	61	2,033333333	1,274712644
Claire	30	61	2,033333333	1,205747126
Ilse ver. Groen	30	71	2,366666667	1,757471264
Xander	30	81	2,7	1,182758621
Controle	30	124	4,133333333	0,947126437

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	160,32	9	17,81333333	14,7511898	1,12E-19	1,912235867
Binnen groepen	350,2	290	1,207586207			
Totaal	510,52	299				

Figuur 4.5: Resultaten van One-way ANOVA in excel voor de lineaire schaalvraag "Hoe natuurlijk vind je deze stem klinken?". De F-waarde is 17.813 en overschrijdt het kritische gebied van F-toets.

Unifactoriële variantie-analyse

SAMENVATTING VERSTAANBAARHEID

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	30	99	3,3	0,906897
Arno	30	83	2,766666667	1,081609
Femke	30	92	3,066666667	1,098851
Ilse ver. Blauw	30	117	3,9	0,437931
Ruben	30	118	3,933333333	0,96092
Guus	30	113	3,766666667	1,633333
Claire	30	111	3,7	0,975862
Ilse ver. Groen	30	119	3,966666667	0,722989
Xander	30	114	3,8	0,855172
Controle	30	121	4,033333333	0,929885

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	49,93666667	9	5,548518519	5,777631	2,15E-07	1,912235867
Binnen groepen	278,5	290	0,960344828			
Totaal	328,4366667	299				

Figuur 4.6: Resultaten van One-way ANOVA in excel voor de lineaire schaalvraag "Hoe verstaanbaar vind je deze stem?". De F-waarde is 5.778 en overschrijdt het kritische gebied van F-toets.

Uit de figuren 4.3 tot en met 4.6 zijn de tabellen van een samenvatting van de groepen op een onderdeel en bijbehorende variantie analyse te zien. De kolom *Aantal* geeft de hoeveelheid waarnemingen van de groep aan. Elk systeem heeft een aantal van 30, dus er zijn 30 proefpersonen geweest die een score hebben toegekend aan een TTS systeem. De kolom *Som* geeft de totale puntentelling op op een desbetreffende onderdeel. De kolom *Variantie* geeft de variantie waardes aan. De variantie geeft de maat aan hoeveel de waardes onderling van elkaar verschillen. De standaarddeviatie is de wortel van de variantie.

In figuur 4.3 heeft het TTS systeem Ilse ver.B de hoogste gemiddelde en Claire het laagste. In figuur 4.4 heeft Ruben de hoogste gemiddelde en Google (RSPV stem) het laagst. In figuur 4.5 heeft de controlestem de hoogste gemiddelde en Google (RSPV stem) het laagst en in figuur 4.6 heeft de controlestem de hoogste gemiddelde en Arno het laagst. In de variantie-analyse tabel in de figuren 4.3 tot en met 4.6 zijn de *kwadratensom*, *vrijheidsgraden*, *gemiddelde kwadraten*, *F*, *P-waarde* en *Kritische gebied van F-toets* te zien. Hierbij wordt er vooral gelet naar de laatste drie kolommen.

De P-waarde is de kans dat de F-waarde door toeval gevonden is. Wanneer de P-waarde kleiner is dan 0.05, dan betekent het dat de kans dat de F-waarde door toeval is gevonden, kleiner is dan 5%. De P-waarde uit elke One-way ANOVA analyse is zeer klein en de F-waarde uit de categorie articulatie, melodie, natuurlijkheid en verstaanbaarheid overtreft de kritische gebied van de F-toets. Dit geeft aan dat het verschil tussen de groepen significant is, maar ook hier wordt er slechts aangegeven dat er een verschil tussen de groepen bestaat en niet waar het verschil ligt. Hiervoor moet er een post hoc analyse gedaan worden.

Er wordt gebruik gemaakt van de Scheffé test om te achterhalen welke tweetal van TTS systemen van elkaar significant verschillen en of de TTS systemen in groepen kan worden geordend die van elkaar verschillen.

$$F_s = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_w^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

Figuur 4.7: Weergave van Scheffé formule om de Scheffé F_s waarde te berekenen.

De F_s-waarde wordt aan de hand van de formule in figuur 4.7 berekent. De S_w^2 is de kwadratensom binnen groepen. n_1 en n_2 zijn het totaal aantal waarnemingen van de desbetreffende groep 1 of 2. De F_s wordt dan vergeleken met de Scheffé kritieke waarde, die verkregen kan worden uit de variantieanalyse door de vrijheidsgraden te vermenigvuldigen met het kritische gebied van de F-Toets. Voor de ANOVA analyse uit de figuren 4.3 tot en met 4.6 komt dan uit dat de Scheffé kritieke waarde uitkomt op 17.210.

Gepaarde TTS systemen	F_s-waarde	Gepaarde TTS systemen	F_s-waarde
Google - Ilse ver.Blauw	24.981	Arno - Ilse ver.Blauw	18.353
Google - Ruben	35.404	Arno - Ruben	27.417
Google - Xander	17.348	Arno - Controle	20.449
Google - Controle	27.417	Ruben - Guus	20.449

Tabel 4.7: F_s-waardes van paren TTS systemen die de Scheffé kritieke waarde overschrijden van de ANOVA analyse op categorie **melodie**.

Gepaarde TTS systemen	F_s-waarde	Gepaarde TTS systemen	F_s-waarde
Google - Ilse ver.Blauw	17.887	Ilse ver. Blauw - Controle	22.083
Google - Ruben	30.488	Guus - Controle	54.779
Google - Controle	79,718	Claire - Controle	54.779
Arno - Ruben	27.948	Ilse ver. Groen - Controle	38.769
Arno - Controle	75.578	Xander - Controle	25.519
Femke - Controle	51.356		

Tabel 4.8: F_s-waardes van paren TTS systemen die de Scheffé kritieke waarde overschrijden van de ANOVA analyse op categorie **natuurlijkheid**.

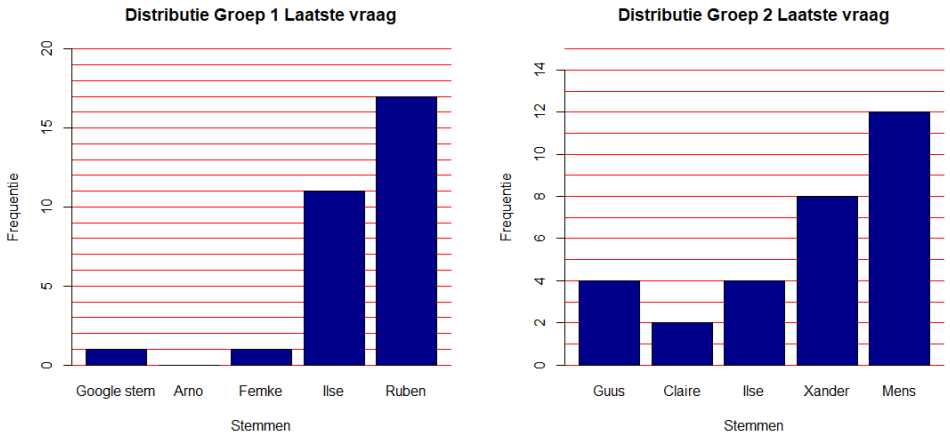
Gepaarde TTS systemen	F_s-waarde	Gepaarde TTS systemen	F_s-waarde
Arno - Ilse ver.Blauw	20.062	Arno - Ilse ver.Groen	22.492
Arno - Ruben	21.260	Arno - Controle	25.060

Tabel 4.9: F_s-waardes van paren TTS systemen die de Scheffé kritieke waarde overschrijden van de ANOVA analyse op categorie **verstaanbaarheid**.

Tabellen 4.7 tot en met 4.9 geven de paren TTS systemen aan waar er onderling significant verschil bestaat. Voor de categorie articulatie was er geen paar van TTS systemen die de Scheffé kritieke waarde overtreffen. Het is mogelijk dat er ondanks dat er significant verschil is geconstateerd uit de ANOVA analyse dat het zwakke effect zo verspreid is tussen de paren TTS systemen dat er geen paar in het bijzonder uitspringt. In tabel 4.7 is er te zien dat de TTS systemen Arno en Google (RSPV stem) het vaakst in paren voorkomen. Dit kan erop duiden dat beide systemen het meest afwijken van de groep TTS systemen bestaande uit Ilse ver. Blauw, Ruben en Controle. Er kan hier mogelijk gesteld worden dat deze systemen erg laag zijn beoordeeld op het onderdeel melodie wat te zien is in tabel 4.3. Dit kan ook gezien worden naar de gemiddelde waardes in figuur 4.4

In tabel 4.8 zien we weer de systemen Google (RSPV stem) en Arno voorkomen in paren, maar opvallend komt de controlestem het vaakst naar voren in paren. Dit duidt erop dat de controlestem het meest afwijkt van alle systemen. In tabel 4.5 is te zien dat de controlestem de hoogste score heeft van alle TTS systemen. Dit duidt erop dat de controlestem zeer goed heeft gepresteerd op deze vraag en voor de systemen Arno en Google (RSPV stem) kan erop duiden dat de systemen op dit onderdeel minder goed hebben gepresteerd op basis van hun scores.

In tabel 4.9 komen alleen de paren naar voren waar Arno voorkomt. Dit kan erop duiden dat het systeem Arno op het onderdeel verstaanbaarheid als laagst werd beoordeeld als er ook gekeken wordt naar de gemiddelde score van het systeem Arno op het onderdeel verstaanbaarheid.



(a) Resultaten vraag beste fragment van enquête Groep 1

(b) Resultaten vraag beste fragment van enquête Groep 2

Figuur 4.8: Resultaten op de vraag: "Welke fragment vond je het best?"van proefpersonen uit groep 1 en groep 2. De als best ervaren fragmenten werden weer herleid naar de TTS systeem. Dit kan gezien worden als de beoordeling van de systemen op algemene prestatie.

Uit de grafieken in figuur 4.2 is er te zien dat vooral de kunstmatige stem Ruben en de menselijke controlestem als beste ervaren zijn. Uit de tabel 4.1 en 4.2 kan er ook gezien worden dat deze twee stemmen relatief hoog beoordeeld waren. De proefpersonen hebben vaker een mannelijke stem gekozen als het beste fragment van alle fragmenten binnen één enquête. In deze grafiek is er ook te zien dat het aantal fragmenten van TTS systeem Ilse minder vaak als beste wordt gekozen in groep 2 dan in groep 1.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,860
Bartlett's Test of Sphericity	Approx. Chi-Square	974,143
	df	21
	Sig.	0,000

Figuur 4.9: Kaiser-Meyer-Olkin (KMO) en Bartlett's test uitslag van de verkregen resultaten.

Component Matrix[a]							
	Component						
	1	2	3	4	5	6	7
Bruikbaarheid	0,839	-0,158	0,009	0,138	-0,438	0,101	-0,225
Melodie	0,814	-0,352	-0,101	-0,215	-0,138	-0,113	0,353
Verstaanbaarheid	0,774	0,416	-0,209	-0,098	0,036	-0,397	-0,124
Natuurlijkheid	0,769	-0,338	-0,233	-0,280	0,313	0,179	-0,181
Articulatie	0,710	0,577	-0,173	0,085	0,013	0,318	0,156
Informatieopname	0,709	-0,201	0,110	0,622	0,220	-0,082	0,046
Passendheid	0,657	0,137	0,700	-0,232	0,069	0,015	-0,012
Extraction Method: Principal Component Analysis.							
a. 7 components extracted.							

Figuur 4.10: Componentenmatrix met de waardes van verschillende variabelen.

Correlation Matrix								
		Verstaanbaarheid	Articulatie	Melodie	Natuurlijkheid	Bruikbaarheid	Informatieopname	Passendheid
Correlation	Verstaanbaarheid	1,000	0,672	0,522	0,493	0,540	0,416	0,440
	Articulatie	0,672	1,000	0,391	0,400	0,506	0,406	0,408
	Melodie	0,522	0,391	1,000	0,701	0,677	0,499	0,450
	Natuurlijkheid	0,493	0,400	0,701	1,000	0,579	0,459	0,387
	Bruikbaarheid	0,540	0,506	0,677	0,579	1,000	0,599	0,478
	Informatieopname	0,416	0,406	0,499	0,459	0,599	1,000	0,385
	Passendheid	0,440	0,408	0,450	0,387	0,478	0,385	1,000

Figuur 4.11: De correlatie matrix van 7 variabelen, afgeleid van de gestelde vragen in een luisterexperiment.

In figuur 4.9 zijn de waardes te zien van de Kaiser-Meyer-Olkin (KMO) en Bartlett's test of Sphericity. Een KMO test is een maat om aan te geven of de verkregen dataset geschikt is voor factor analyse. Wanneer deze waarde dichtbij de waarde 1 komt, betekent dat een factor analyse bruikbaar is voor de dataset. Een KMO waarde onder de 0.6 geeft aan dat factor analyse niet geschikt is voor een dataset (Cerny et Kaiser, 1977) [28]. Uit de KMO test komt er een Kaiser-Meyer-Olkin Measure of Sampling Adequacy waarde van 0.860 uit. De Bartlett's test of Sphericity geeft aan of de correlatiematrix gelijk is aan een identiteitsmatrix. Dit houdt in dat de test kan aantonen of er correlaties bestaan tussen verschillende variabelen. Wanneer een significantiewaarde boven 0.05 wordt bereikt, betekent dit dat de variabelen niet met elkaar gecorreleerd zijn en ook niet bruikbaar zijn voor factor analyse. Aangezien er een significantiewaarde lager dan 0.05 is behaald, kan er aangenomen worden dat er correlaties tussen variabelen bestaan (Tobias et Carlson, 1969) [29].

In figuur 4.4 is er een componentenmatrix te zien waar elke rij een de naam van een variabele bevat en elke kolom een principale component. De Principal component analysis is een analyse methode dat ervoor zorgt dat een grote dataset met veel variabelen verklaard kan worden aan de hand van een kleinere set variabelen. Bij principale component 1 hebben de variabelen een relatief hoge waardes. De variabelen in hoofdcomponent 1 zijn met elkaar gecorreleerd. Dit kan betekenen dat wanneer de waarde bij één van de variabelen stijgt, dat de waardes van andere variabelen ook stijgen. De waarde van de variabele bruikbaarheid is daarentegen bij hoofdcomponent 1 het hoogst. De eerste component kan worden beschreven als "Totale prestatie van een TTS systeem". Voor principale component 2 hebben de variabelen articulatie en verstaanbaarheid een relatief hoge waarde. Dit kan mogelijk beschreven worden als "De helderheid van spraak van een TTS systeem". In component 3 is de variabele passendheid als enige die een relatieve hoge waarde heeft en in component 4 is dat informatieopname.

In figuur 4.5 kunnen de correlatiecoëffcient tussen verschillende variabelen gezien worden. Deze waardes geven aan hoe sterk het verband is tussen de variabelen. Een correlatiecoëffcient met de waarde 0 geeft aan dat er geen correlatie is tussen twee variabelen. Hoe verder de waarde van de 0 verwijderd is, des te sterker is de correlatie tussen de variabelen (Chu et al., 2013) [30]. Er is te zien dat de correlatiecoëffcient tussen natuurlijkheid en melodie 0.701 is. Dit is een relatief hoge waarde en kan dus betekenen dat er een verband bestaat. Dit geldt ook tussen de variabelen articulatie en verstaanbaarheid.

Met de factorladingen die in de componententmatrix in figuur 4.10 staan kunnen de scores die proefpersonen hebben toegekend aan TTS systemen omgerekend worden naar een totale score per proefpersoon voor een TTS systeem in dimensie 1 tot en met 4. De totale scores van proefpersonen op TTS systemen per dimensie kan dan ook geanalyseerd worden met One-way ANOVA.

Unifactoriële variantie-analyse

SAMENVATTING DIMENSIE 1

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	25	231,989	9,27956	6,421748528
Arno	25	219,154	8,76616	8,841390703
Femke	25	257,841	10,31364	12,74081097
Ilse ver. Blauw	25	343,143	13,72572	10,99220611
Ruben	25	360,1095	14,40438	15,63363213
Guus	25	284,35	11,374	14,71001456
Claire	25	277,209	11,08836	14,99328143
Ilse ver. Groen	25	310,6105	12,42442	11,68446151
Xander	25	329,514	13,18056	10,34041905
Controle	25	345,334	13,81336	15,62434114

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	882,023302	9	98,00258911	8,034164316	2,16E-10	1,919026242
Binnen groepen	2927,575347	240	12,19823061			
Totaal	3809,598649	249				

Figuur 4.12: Resultaten van One-way ANOVA in excel voor hoofdcomponent 1. De F-waarde is 8.034 en overschrijdt het kritische gebied van F-toets.

Unifactoriële variantie-analyse

SAMENVATTING DIMENSIE 2

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	25	-23,045	-0,9218	0,495616
Arno	25	-11,966	-0,47864	0,536802
Femke	25	-2,873	-0,11492	0,774498
Ilse ver. Blauw	25	-7,118	-0,28472	0,686758
Ruben	25	-0,3405	-0,01362	0,587607
Guus	25	-23,649	-0,94596	0,803531
Claire	25	-20,285	-0,8114	0,752714
Ilse ver. Groen	25	-15,1025	-0,6041	0,956642
Xander	25	-4,523	-0,18092	0,45637
Controle	25	17,65	0,706	0,743934

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	57,2568954	9	6,361877267	9,363317	3,5E-12	1,919026242
Binnen groepen	163,0672686	240	0,679446952			
Totaal	220,324164	249				

Figuur 4.13: Resultaten van One-way ANOVA in excel voor hoofdcomponent 2. De F-waarde is 9,363 en overschrijdt het kritische gebied van F-toets.

Unifactoriële variantie-analyse

SAMENVATTING DIMENSIE 3

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	25	-36,983	-1,47932	0,090418
Arno	25	-35,708	-1,42832	0,231831
Femke	25	-39,083	-1,56332	0,251386
Ilse ver. Blauw	25	-46,297	-1,85188	0,142018
Ruben	25	-48,994	-1,95976	0,225991
Guus	25	-42,361	-1,69444	0,269634
Claire	25	-39,281	-1,57124	0,191267
Ilse ver. Groen	25	-46,214	-1,84856	0,194102
Xander	25	-46,397	-1,85588	0,18995
Controle	25	-53,036	-2,12144	0,233855

Variantie-analyse

Bron van variatie	Kwadratensom	Vrijheidsgraden	Gemiddelde kwadraten	F	P-waarde	Kritische gebied van F-toets
Tussen groepen	11,45811754	9	1,273124171	6,30119	5,3E-08	1,919026242
Binnen groepen	48,4908092	240	0,202045038			
Totaal	59,94892674	249				

Figuur 4.14: Resultaten van One-way ANOVA in excel voor hoofdcomponent 3. De F-waarde is 6.301 en overschrijdt het kritische gebied van F-toets.

Unifactoriële variantie-analyse

SAMENVATTING DIMENSIE 4

Groepen	*Aantal*	*Som*	*Gemiddelde*	*Variantie*
Google	25	-12,012	-0,48048	0,446811
Arno	25	-8,978	-0,35912	0,412665
Femke	25	-4,784	-0,19136	0,303772
Ilse ver. Blauw	25	-0,631	-0,02524	0,411616
Ruben	25	2,701	0,10804	0,57521
Guus	25	-13,317	-0,53268	0,321252
Claire	25	-9,754	-0,39016	0,347032
Ilse ver. Groen	25	-7,672	-0,30688	0,679462
Xander	25	-5,54	-0,2216	0,380026
Controle	25	-5,24	-0,2096	0,317371

Variantie-analyse

Bron van variatie	*Kwadratensom*	*Vrijheidsgraden*	*Gemiddelde kwadraten*	*F*	*P-waarde*	*Kritische gebied van F-toets*
Tussen groepen	8,780339284	9	0,975593254	2,325489	0,015837	1,919026242
Binnen groepen	100,6852136	240	0,419521723			
Totaal	109,4655529	249				

Figuur 4.15: Resultaten van One-way ANOVA in excel voor hoofdcomponent 4. De F-waarde is 2.325 en overschrijdt het kritische gebied van F-toets.

In figuren 4.12 tot en met 4.15 is er ook te zien dat de F-waarde het kritische gebied van F-toets overschrijdt. Hiermee kan dus ook gesteld worden dat er tussen de groepen van dimensie 1 tot en met 4 een significant verschil bestaat. Er moet hier ook een post hoc analyse gedaan worden om te achterhalen waar dit verschil precies ligt. De Scheffé methode wordt weer toegepast op elke dimensie. Het nieuwe kritische gebied van F-toets voor Scheffé methode is 17.271.

Gepaarde TTS systemen	F_s-waarde	Gepaarde TTS systemen	F_s-waarde
Google - Ilse ver.Blauw	20.257	Arno - Ruben	32.576
Google - Controle	21.063	Arno - Xander	19.969
Arno - Ilse ver.Blauw	25.207	Arno - Controle	26.104

Tabel 4.10: F_s-waardes van paren TTS systemen die de Scheffé kritieke waarde overschrijden van de ANOVA analyse van Dimensie 1.

Gepaarde TTS systemen	F_s-waarde	Gepaarde TTS systemen	F_s-waarde
Google - Controle	48.748	Guus - Controle	50.206
Arno - Controle	25.818	Claire - Controle	42.360
Ilse ver. Blauw - Controle	18.057	Ilse ver. Groen - Controle	31.576

Tabel 4.11: F_s-waardes van paren TTS systemen die de Scheffé kritieke waarde overschrijden van de ANOVA analyse van Dimensie 2.

Gepaarde TTS systemen	F_s-waarde	Gepaarde TTS systemen	F_s-waarde
Google - Controle	25.509		

Tabel 4.12: F_s-waardes van paren TTS systemen die de Scheffé kritieke waarde overschrijden van de ANOVA analyse van Dimensie 3.

De resultaten van de Scheffé methode zijn weergegeven in tabellen 4.10, 4.11 en 4.12. In tabel 4.10 komt het systeem Arno het meest voor in de paren. Eerder werd de eerste component beschreven als de "Totale prestatie van TTS systeem". Er kan hier gesteld worden dat het systeem Arno lager heeft gepresteerd ten opzichte van de systemen Ilse van groep 1, Ruben en Xander.

In tabel 4.11 komt de controlestem in alle paren voor. Hoofdcomponent 2 werd beschreven als "De helderheid van spraak van TTS systeem". De controlestem wijkt het meest af van een groot deel van TTS systemen. In figuur 4.13 kan gezien worden dat de controlestem een hoge gemiddelde heeft ten opzicht van de rest. Dit duidt erop dat de spraak van de controlestem zeer helder is. Voor dimensie 4 zijn er geen paren voorgekomen die het kritische gebied van F-toets overschrijdt. Hier is het ook mogelijk dat het zwakke effect zeer verspreid is dat er geen bijzondere tweetal van TTS systemen eruit springt.

5 Discussie

Voordat er een aantal conclusies getrokken worden, moeten er een aantal zaken vooraf gemeld worden. Tijdens het onderzoek zijn er een aantal zaken naar voren gekomen die impact zouden hebben op de conclusies die getrokken kunnen worden uit de resultaten.

Één van die zaken was dat niet alle stemmen optimaal bleken te zijn bij het voorlezen van teksten. Tijdens het maken van de fragmenten, was er een opmerkelijke probleem gevonden bij de stem van ResponsiveVoice. Tijdens het voorlezen maakte deze stem na 3 tot 4 zinnen een lange pauze van 10 seconden. Deze pauzes zorgden ervoor dat de melodie verbroken wordt en dat de zinnen een vreemde overgang aan prosodie verkrijgen. Er werd eerder in sectie Resultaten vastgesteld dat de stem van ResponsiveVoice het laagst heeft gepresteerd op melodie. Er werd gepoogd deze pauzes zoveel mogelijk te verkorten of eruit te halen. Uiteindelijk zijn de pauzes korter dan wat ze oorspronkelijk waren, maar zijn inmiddels nog steeds niet ideaal. Dit houdt in dat deze factor hoogstwaarschijnlijk invloed heeft gehad op de beoordeling van proefpersonen.

De proefpersonen konden maximaal één keer een audio fragment beluisteren. Dit houdt in dat TTS systemen beoordeeld zijn op een eerste indruk in plaats van gebruik op langere termijn. Het kan mogelijk zijn dat de resultaten die in dit onderzoek zijn verkregen niet overeenkomen met resultaten waarbij TTS systemen wel vaker beluisterd mogen worden. Daarentegen viel de toelichting op de vraag: "Waarom vind je deze stem wel/niet passen bij de tekst?" van sommige proefpersonen op dat zij op langere termijn een TTS systeem hebben gebruikt. Het zou dus eventueel handiger zijn geweest om vooraf een vraag te stellen of een proefpersoon voor persoonlijke doeleindes een TTS systeem zou hebben gebruikt.

Voor een proefpersoon die nooit of lang geleden een kunstmatige stem heeft gehoord kan de beoordeling beïnvloed worden door de eerste TTS systeem die deze persoon hoort. De eerste kunstmatige stem kan voor een proefpersoon een referentie zijn. Dit kan mogelijk invloed hebben op de beoordeling van andere kunstmatige stemmen. Dit kan bij dit onderzoek niet onderzocht worden aangezien er te weinig proefpersonen waren om hier een statistische analyse te doen.

Verder is het voor de groep die wel eerder een kunstmatige stem hebben gehoord ook niet duidelijk op welke manier deze groep mensen in contact zijn gekomen met een kunstmatige stem en in welke maten zij in contact zijn gekomen.

In de resultaten van King et al. (2017) [18] is te gezien dat verschillende TTS systemen die gebruik maken van verschillende technieken significant verschillen op verscheidene onderdelen. Voor een groot deel van de TTS systemen in dit onderzoek is er geen significant verschil aangetoond, op de systemen Arno en ResponsiveVoice na. Het zou voor een mogelijk vervolgonderzoek gunstig zijn als er TTS systemen worden onderzocht die andere technieken gebruikten zoals formant synthese of spraakproductie met neurale netwerken om deze vermoeden te ondersteunen. In dit onderzoek waren er voornamelijk de twee genoemde TTS systemen die significant verschillen van de andere systemen.

Er zijn twee groepen proefpersonen gebruikt in dit onderzoek. Voor het uitvoeren van de statistische analyses werd er vanuit gegaan dat beide groepen gelijkwaardige beoordelingen geven, maar de verschillen van de TTS systeem Ilse in groep 1 en groep 2, kan erop duiden dat dit niet het geval is. Verdere analyse over het impact van deze analyse valt buiten het onderzoek.

Als controlestem werd een menselijke mannenstem gebruikt in het onderzoek. Er

bestaat hier een vermoeden dat de menselijke controlestem invloed zou hebben gehad op de beoordeling van andere TTS mannenstemmen. Om dit na te gaan zou er bij een ander onderzoek met een vrouwelijke controlestem getest moeten worden om dit te achterhalen.

Verder zou het handig zijn geweest dat beide groepen een controlestem bevatten. Dit was niet uitgevoerd, omdat de controlestem op een later moment van het onderzoek verkregen was.

In de theoretische achtergrond werd er besproken dat een simpele implementatie van de techniek unit selection kan zorgen dat de spraaksegmenten niet goed op elkaar gaan aansluiten wat kan leiden naar hakkelende spraak. Dit kan teruggevonden worden in het TTS systeem Arno waar de gemiddelde score voor het onderdeel melodie zeer laag is. Dit kan ook gezien worden in King and Karaiskos (2016) [31] waar drie unit selection TTS systemen meededen bij The Blizzard Challenge 2016 en in King et al. (2017) waar één unit selection TTS systeem mee deed. In de resultaten van The Blizzard Challenge 2016 kan er gezien worden dat deze drie systemen significant slechter presteerden dan de unit selection benchmark TTS systeem op het onderdeel melodie. Deze systemen hadden op een gemiddelde score tussen de 20 en 25 van de 60 punten, wat omgerekend rond een gemiddelde MOS score van 2 ligt. Voor natuurlijkheid lag de score van de systemen gemiddeld tussen de 2 of de 3. Op verstaanbaarheid hadden deze systemen een Word Error rate tussen de 44% en 55%, dus bijna 1 op de twee woorden wordt door proefpersonen verkeerd gehoord. Dit kan niet één-op-één worden getrokken, maar vermoedelijk zou de score op verstaanbaarheid relatief laag liggen. Het TTS systeem in King et al. (2017) had voor melodie gemiddeld 24 van de 60 punten behaald. Voor natuurlijkheid scoorde dit systeem gemiddeld rond de 3. De score van natuurlijkheid van dit systeem komt overeen met de meeste TTS systemen op Arno en ResponsiveVoice na.

Zowel de menselijke controlestem als de controlestem van The Blizzard Challenge hebben een relatief hoge score of de hoogste score op elk onderdeel gehaald. Wel moet er gemeld worden dat er bij de spraaksegmenten van de controlestem voor dit onderzoek duidelijk te horen was dat deze persoon bij s-klanken slist. Dit kan voor veel mensen niet aangenaam klinken, wat mogelijk kan resulteren dat dit invloed gaf op de beoordeling van proefpersonen. Er bestaat een vermoeden dat de beoordeling over de controlestem op de onderdelen articulatie en melodie beínvloed zijn, aangezien de controlestem op deze onderdelen niet als hoogst heeft gescoord, terwijl dit wel het geval is bij de controlestem van The Blizzard Challenge 2017 [18].

Een ander zaak is dat het aantal vragen dat bij een enquête gesteld werd mogelijk te weinig was voor een subjectieve evaluatie. Normaliter wordt meer evaluatiemateriaal gebruikt en vragen gesteld in een MOS test. Dit is terug te zien in de evaluatie test van The Blizzard Challenge. De resultaten die hier verkregen kunnen een indicatie geven, maar zullen niet een volledige representatie kunnen verantwoorden.

6 Conclusie

6.1 Beantwoording onderzoeksvragen

6.1.1 Beantwoording deelvragen

Hoe wordt de natuurlijkheid en verstaanbaarheid van spraak beoordeeld? Uit de resultaten is gebleken dat de scores op de natuurlijkheid van spraak van alle TTS systemen lager zijn dan de score van de controlestem. Dit wordt ook ondersteund door de Scheffé test resultaten uit tabel 4.8 waar de controlestem met elk ander systeem significant verschilt en uit de resultaten van de One-way ANOVA met scheffé test in figuur 4.13 en tabel 4.11. Dit duidt erop dat de kunstmatige stemmen nog niet het niveau van natuurlijkheid bereikt hebben als menselijke spraak en dat synthetische spraak door gebruikers nog als te kunstmatig wordt ervaren. Dit blijkt ook zo te zijn in de resultaten van The Blizzard Challenge 2016 en 2017 waar ook besproken werd dat geen enkele TTS systeem net zo natuurlijk klinkt als menselijke spraak [31] [18]. Voor de verstaanbaarheid van spraak heeft het TTS systeem Arno het laagst gepresteerd. De gemiddelde score van Arno op verstaanbaarheid is dan ook het laagste van alle systemen. Voor de andere systemen werd er bij de Scheffé test onderling geen significante verschil gevonden. Dit kan erop wijzen dat deze TTS systemen voldoende presteren op verstaanbaarheid dat het in praktijk kan worden ingezet.

Worden alle stemmen even goed beoordeeld? Volgens de resultaten is gebleken dat alle stemmen niet even goed worden beoordeeld. Er zijn duidelijk verschillen te zien tussen de scores van groep 1 en groep 2 uit de tabellen 4.1 en 4.3. Volgens de analyse van de sectie Resultaten blijkt dat er een significant verschil bestaat tussen de TTS systemen. Dit duidt erop dat de kunstmatige stemmen van elkaar verschillen en dat de luisteraars dit mogelijk hebben ervaren op basis van de scores die zij hebben gegeven. De kunstmatige stem Arno heeft relatief gezien lagere scores op elke onderdeel ten opzichte van de kunstmatige stem Ruben of Ilse van groep 1 en dit wordt ook ondersteund door de chi-kwadraattoets in paren en Scheffé test analyse. Dit kan ook gelden voor het TTS systeem van ResponsiveVoice als er wordt bekeken naar de gemiddelde scores en de tabellen 4.2, 4.7, 4.8 en 4.10. De TTS systemen Arno en ResponsiveVoice hebben over het geheel het laagst gepresteerd bij dit onderzoek ten opzichte van andere TTS systemen. Er kan gesuggereerd worden dat er mogelijk 3 groepen zijn waar de TTS systemen ingedeeld kan worden. De eerste groep die het laagst heeft gepresteerd, waar Arno en ResponsiveVoice onder vallen. Een tweede groep die onbeslist tussen groep 1 en 3 zweeft, waar Femke, Guus, Claire en Ilse van groep 2 onder vallen en de derde groep die relatief goed heeft gepresteerd waar Ilse van groep 1, Ruben, Xander en controlestem onder vallen.

Wat zijn de belangrijkste verbeterpunten van Nederlandstalige Text-To-Speech? Het belangrijkste verbeterpunt van Nederlandstalige Text-To-Speech is de natuurlijkheid van spraak. Dit blijkt uit de analyse van de resultaten, zoals hiervoor leesbaar is geweest, waar bleek dat alle TTS systemen significant lager scoren dan de controlestem op natuurlijkheid en dat de verschillen in gemiddelde scores ook significant zijn. Ook is er gebleken dat er tussen de variabele melodie en natuurlijkheid een correlatie bestaat volgens de PCA analyse. Dit kan betekenen dat wanneer melodie van spraak wordt verbeterd, dat

het waarschijnlijk is dat natuurlijkheid van spraak ook wordt verbeterd. Aangezien de meeste TTS systemen gebruik maken van de techniek unit selection, sluit dit aan bij de theorie.

Kan Nederlandstalige Text-To-Speech voor meer toepassingen worden ingezet als er verbeteringen worden doorgevoerd? Nederlandstalige Text-To-Speech wordt op dit moment gebruikt op bijvoorbeeld navigatiesystemen of smartphones en biedt ondersteuning aan mensen die visueel gehandicapt zijn. Op de site van Readspeaker staat enkele toepassingen waar TTS systemen ingezet kan worden. Net binnengekomen nieuwsartikelen kunnen gelijk van spraak worden voorzien of websites kunnen veel toegankelijker gemaakt worden voor ouderen en visueel gehandicapten.

Het kan eventueel ook gebruikt worden voor huishoudelijke apparaten. In een bijeenkomst van november 2014 van NoTaS (Nederlandse Organisatie voor Taal- en Spraaktechnologie) was het wenselijk dat de bediening en terugmelding van bijvoorbeeld een wasmachine of kookplaat doormiddel van spraak zou gebeuren (Hessen et al.,2018) [32].

Het zou ook mogelijk zijn om robots van spraak te voorzien om mogelijk kleine taken over te nemen zoals wat er in Japan gebeurt. In Japan is er een robot genaamd Pepper die werk verricht als boeddhistische priester. Pepper is voorgeprogrammeerd om lange sutra's uit te spreken en kan dus ook uitvaartdiensten uitvoeren (Cheh, 2017) [33]. Text-To-Speech kan dus op veel fronten worden ingezet.

6.1.2 Beantwoording hoofdvraag

Wat is de huidige kwaliteit van Nederlandse spraaksynthese? Uit de resultaten en deelvragen kan er opgemaakt worden dat de TTS systemen in groepen verdeeld kunnen worden op basis van kwaliteit. De TTS systemen Arno en ResponsiveVoice hebben op basis van de beoordeling van proefpersonen voor dit onderzoek het laagst gepresteerd ten opzichte van andere TTS systemen. Als geheel wordt de kwaliteit van Nederlandse spraaksynthese nog niet beoordeeld als de menselijke controlestem. Er kan wel gesteld worden dat er een verschil in kwaliteit bestaat tussen TTS systemen, ondanks dat bijna alle systemen de techniek unit selection gebruiken. Dit kan betekenen dat het mogelijk is om de kwaliteit van TTS systemen zodanig te verbeteren dat het door eindgebruikers als prettig wordt ervaren en dat meer onderzoek naar TTS systemen kan lonen. Dit kan gezien worden als bijvoorbeeld het TTS systeem Arno wordt vergeleken met Ruben in de sectie resultaten. Dit kan ook gezien worden wanneer er gekeken wordt naar de resultaten van de unit selection TTS systemen van The Blizzard Challenge 2016 (King and Karaiskos, 2016) en van The Blizzard Challenge 2017 (2017) wat besproken werd in de discussie.

6.2 Conclusie

In dit onderzoek werd er gezocht naar een antwoord op de vraag: "Wat is de huidige kwaliteit van Nederlandse spraaksynthese?". Hiervoor werd een subjectieve evaluatie in een vorm van een enquête uitgevoerd waarbij in totaal 60 studenten meededen.

Uit de resultaten blijkt dat sommige TTS systemen beter beoordeeld waren dan andere systemen door proefpersonen en dat de systemen mogelijk in 3 groepen gedeeld kan worden aan de hand van de resultaten van de analyse. De verstaanbaarheid van TTS systemen scoort gemiddeld tussen de 3 en de 4, wat aangeeft dat TTS systemen voldoende verstaanbaar zijn om ingezet te worden. De gemiddelde scores van de natuurlijkheid van synthetische spraak ligt daarentegen tussen de 2 en 3. De scores van natuurlijkheid komen overeen in de resultaten van de unit selection systemen van The Blizzard Challenge 2016 en 2017 [31] [18]. Dit is dan ook één van de belangrijkste verbeterpunten om de kwaliteit

van Nederlandse spraaksynthese te verbeteren samen met de melodie. Er is ook gebleken dat er tussen de melodie en de natuurlijkheid een verband bestaat volgens de PCA analyse. Wanneer de melodie wordt verbeterd, dan is het mogelijk dat ook de natuurlijkheid van de synthetische spraak beter wordt beoordeeld. Voor vervolgonderzoek is het gewenst om TTS systemen te betrekken die andere technieken gebruiken en in het bijzonder met de techniek DNN, aangezien er in de Theoretische kader eerder werd besproken dat het een opkomende techniek is waar veel potentie in kan zitten voor het produceren van synthetische spraak en dat de verbeterpunten die in de discussie werden besproken ook meegenomen moeten worden voor een vervolgonderzoek.

Bijlages

A Experiment teksten

Tekst 1: Nu nieuwsartikel 'Mensen met een beperking ondervinden veel problemen op gebied van werk' 1-12-2017
Nederland doet te weinig om mensen met een beperking volledig te laten deelnemen aan de samenleving. Naast problemen bij zelfstandig wonen, toegankelijkheid en onderwijs, zouden zij vooral veel last ondervinden op het gebied van arbeid. Dat concludeert het College voor de Rechten van de Mens in een vrijdag verschenen rapport. Het rapport is opgesteld naar aanleiding van het VN-verdrag voor rechten van mensen met een handicap, dat ruim een jaar geleden is ingegaan. Het wordt vrijdag overhandigd aan minister Hugo de Jonge (Volksgezondheid).

Tekst 2: Interview Scientias jan Broersen Universiteit Utrecht 22-10-2017
Bij kunstmatige intelligentie denken veel mensen al snel aan robots die al dan niet het slechtste met de mensheid voor hebben. Maar kunstmatige intelligentie is veel breder dan dat en het is zelfs zo breed dat er niet echt een eenduidige definitie voor bestaat. Dat komt mede doordat ?intelligentie? zelf zo lastig te vangen is in een definitie. Jan Broersen, universitair hoofddocent en onderzoeker aan de Universiteit Utrecht, vat het samen als ?het via computationele middelen proberen nabootsen van onze intelligentie.

Tekst 3: Onderzoeksgids-Geschiedenis Jacco Pekelder 1-1-2015
Archiefonderzoek. Bronnenonderzoek is een van de leukste en spannendste vormen van historisch onderzoek. Het biedt de mogelijkheid om interessante vondsten te doen en kan je in staat stellen de bestaande wetenschappelijke literatuur over een bepaald thema aan te vullen of te corrigeren. Houd er rekening mee dat een archiefvormer ook niet neutraal is. Een archief kan ook gebruikt worden om een wenselijke geschiedenis achter te laten. Als je ervoor kiest om voor een historisch werkstuk of scriptie archiefonderzoek te doen, dan moet je wel beseffen dat enige voorkennis vereist is. Je moet vertrouwd raken met het archiefwezen.

Tekst 4 Inleiding in de Wiskunde S.A. Terwijn
3.3 Equivalentierelaties. Een essentieel aspect van de wiskunde is abstractie, dat wil zeggen, het weglaten van onnodige details en eigenschappen van het onderwerp van studie. Hoewel bijvoorbeeld lijnen in de praktijk altijd een bepaalde dikte hebben, hebben de lijnen in de Euclidische meetkunde helemaal geen dikte. Grappigerwijze vergroot deze idealisering juist de toepasbaarheid van de theorie. Of we kunnen bijvoorbeeld in bepaalde situaties waarin alleen telt op welke dag iemand geboren is redeneren over mensen, waarbij we personen met dezelfde verjaardag als gelijk beschouwen. Deze laatste situatie, waarin we groepen objecten indelen volgens een bepaald criterium, komt in de wiskunde veelvuldig voor, en wordt geformaliseerd door het begrip equivalentierelatie.

Tekst 5: Nu nieuwsartikel Leraren basisonderwijs gaan dinsdag 12 december weer staken 5-12-2017
De leraren in het primair onderwijs gaan op dinsdag 12 december weer staken. Dat heeft actiegroep PO in actie bekendgemaakt nadat een ultimatum aan het kabinet voor het vrijmaken van extra geld was verlopen. De ontevreden leerkrachten, onder leiding van PO

in actie, willen 1,4 miljard euro voor het basisonderwijs. Het geld, 900 miljoen euro voor een salarisverhoging en 500 miljoen voor het verlichten van de werkdruk, is bedoeld om het dreigende lerarentekort af te wenden. Maar minister Arie Slob (Onderwijs) zegt niet meer te kunnen bieden dan de driekwart miljard die in het regeerakkoord is afgesproken.

B Opzet Testversies

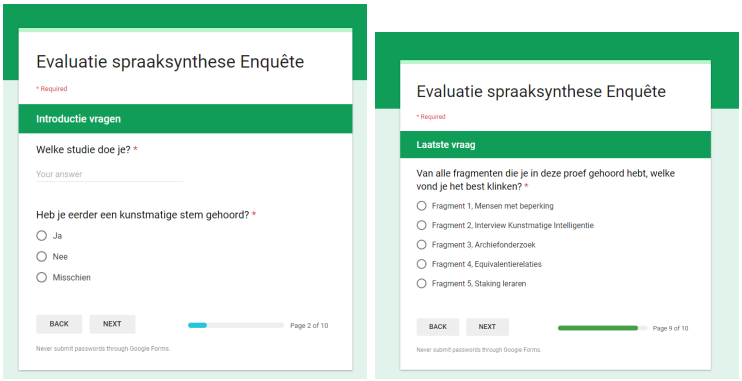
Versie	Tekst 1	Tekst 2	Tekst 3	Tekst 4	Tekst 5
1	RSPV stem	Arno	Femke	Ilse	Ruben
2	Arno	Femke	Ilse	Ruben	RSPV stem
3	Femke	Ilse	Ruben	RSPV stem	Arno
4	Ilse	Ruben	RSPV stem	Arno	Femke
5	Ruben	RSPV stem	Arno	Femke	Ilse

Tabel B.1: Groep 1 enquêteversies

Versie	Tekst 1	Tekst 2	Tekst 3	Tekst 4	Tekst 5
1	Guus	Claire	Ilse	Xander	Controlestem
2	Claire	Ilse	Xander	Controlestem	Guus
3	Ilse	Xander	Controlestem	Guus	Claire
4	Xander	Controlestem	Guus	Claire	Ilse
5	Controlestem	Guus	Claire	Ilse	Xander

Tabel B.2: Groep 2 enquêteversies

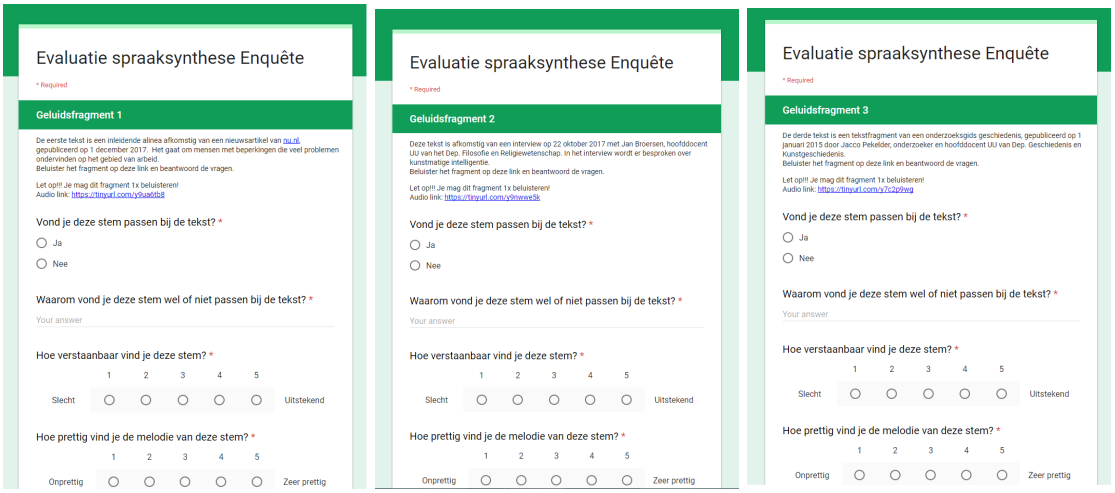
C Screenshots Enquete



(a) Eerste vraagsectie van enquête.

(b) Laatste vraagsectie van enquête.

Figuur C.1: Begin en eindpagina's van een enquête

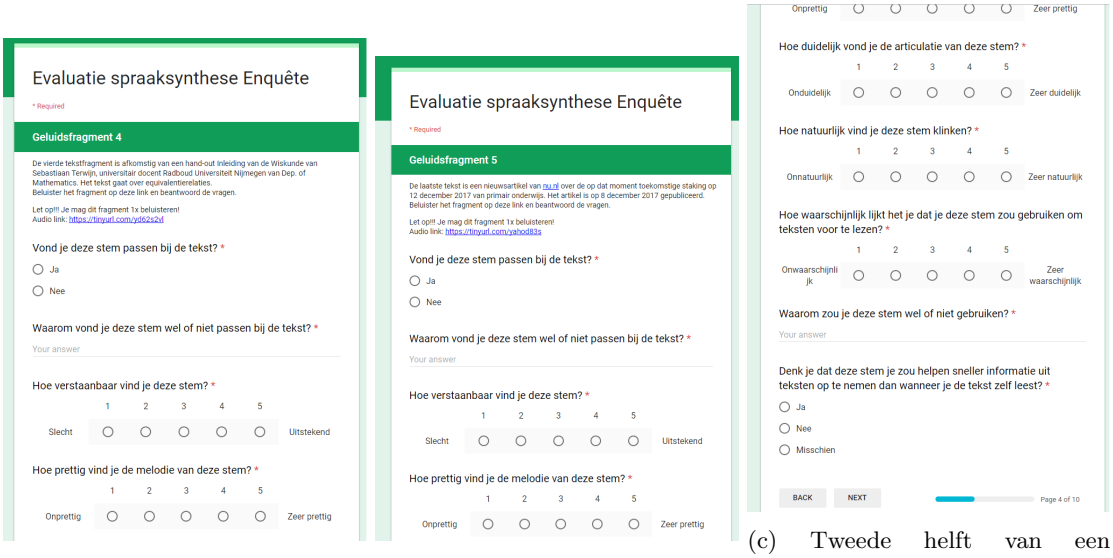


(a) Fragment 1 vragensectie van enquete.

(b) Fragment 2 vragensectie van enquete.

(c) Fragment 3 vragensectie van enquete

Figuur C.2: Google Form enquête



(a) Fragment 4 vragensectie van enquete.

(b) Fragment 5 vragensectie van enquete.

(c) Tweede helft van een enquête.(Voor elke test hetzelfde)

Figuur C.3: Google Form enquête

D Overige resultaten enquête

	Studies	Freq
1	Aardwetenschappen	3
2	Archeologie	1
3	Biologie	1
4	Biologie en Laboratoriumonderzoek	1
5	Bouwkunde	1
6	Climate Physics	1
7	Computer Science and Engineering	2
8	Computing Science	2
9	Diergeneeskunde	2
10	Econometrie	1
11	Electrical Engineering	2
12	Environmental Studies	1
13	Farmaceutische wetenschappen	1
14	financieel recht en economie	1
15	Fine arts	1
16	Fiscaal Recht	1
17	Game and Media technologie (uu informatica master)	1

	Studies	Freq
18	geschiedenisleraar	1
19	HBO-ICT	1
20	HBO-ICT-SIE (Software & Information Engineering)	1
21	HBO Life Sciences	1
22	HBO Natuurkunde	1
23	HBO PABO	1
24	Human Resource Management	1
25	Industrieel ontwerpen, TU Delft	1
26	Informatica	2
27	Information & Media Studies	1
28	Integrale Veiligheidskunde	1
29	Japanstudies	1
30	Keltische Talen en Cultuur	2
31	Kopopleiding docent Omgangskunde	1
32	Kunstmatige Intelligentie	4
33	Mechatronica	1
34	Medische Hulpverlening	1

	Studies	Freq
34	Medische Hulpverlening	1
35	Milieukunde	1
36	Mondzorgkunde	1
37	Natuurkunde	1
38	Pharmaceutical sciences	1
39	Product Design	1
40	Psychologie	1
41	Scheikunde	2
42	Software Science	2
43	Technische Informatica	1
44	Technische Natuurkunde	2
45	Werktuigbouwkunde	1
46	Wiskunde	1

Figuur D.1: Frequentie van studies

E Chi-kwadraat tabel

Degrees of freedom	α									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figuur E.1: Chi-kwadraat tabel van kritieke waardes. Afhankelijk van de α en het aantal vrijheidsgraden, kan de kritieke waarde afgelezen worden.

Bibliografie

[1] S. Hawking, *Stephen Hawking: My Computer*, 2018. adres: `http://www.hawking.org.uk/the-computer.html`.

[2] M. Ruiter, „Taal- en spraaktechnologie en communicatieve beperkingen: Behoeften en mogelijkheden voor de toekomst", *Taalunie*, 2010.

[3] A. C. Rietveld en V. J. Heuven, *Algemene fonetiek*. Coutinho, 2016.

[4] D. Jurafsky en J. H. Martin, *Speech and language processing 2nd edition*. Pearson Education, 2008, p.283-315.

[5] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf en C. Richards, „Normalization of non-standard words", *Computer Speech and Language*, jrg. 15, nr. 3, 2001, p. 287-333. adres: `https://doi.org/10.1006/csla.2001.0169`.

[6] T. Dutoit, „High-quality text-to-speech synthesis : an overview", *Journal of Electrical & Electronics Engineering*, 1997.

[7] M. Bisani en H. Ney, „Joint-Sequence Models for Grapheme-to-Phoneme Conversion Speech Communication", *Elsevier : North-Holland*, 2008.

[8] N. Campbell, „Evaluation of Speech Synthesis", in *Evaluation of Text and Speech Systems*. Dordrecht: Springer Netherlands, 2007, pp. 29–64, ISBN: 978-1-4020-5817-2.

[9] Y. Xu, „Speech Prosody — Theories, models and analysis", in *Courses on Speech Prosody*. Cambridge Scholars Publishing, 06-2015, pp. 146–177.

[10] H. Quené en R. Kager, „EuroSpeech '89", *European Conference on Speech Communication and Technology*, jrg. 1, nr. september, 1989, p. 214 - 217.

[11] S. E. Levinson, „Articulatory speech synthesis from the fluid dynamics of the vocal apparatus", *San Rafael: Morgan & Claypool*, 2012.

[12] E. Klabbers, „Smart Voices", *Dixit: AI en TST*, 2017.

[13] P. Taylor, *Text-to-speech synthesis*. Cambridge University Press, 2009.

[14] A. J. Hunt en A. W. Black, „Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", in *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. On Conference Proceedings., 1996 IEEE International Conference - Volume 01*, reeks ICASSP '96, IEEE Computer Society, 1996, pp. 373–376, ISBN: 0-7803-3192-3.

[15] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior en K. Kavukcuoglu, „WaveNet: A Generative Model for Raw Audio", in *Arxiv*, 2016. adres: `https://arxiv.org/abs/1609.03499`.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier en A. P. Hekstra, „Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part II: Psychoacoustic Model", *Journal of the AES*, nr. 10, 50 2002.

[17] H. Cryer en S. Home, „Review of methods for evaluating synthetic speech", RNIB Centre for Accessible Information (CAI), tech. rap., 02-2010.

[18] S. King, W. Lovisa en W. Guo, „The Blizzard Challenge 2017", *The Centre for Speech Technology Research University of Edinburgh*, 2017.

[19] M. Cernak en M. Rusko, „An evaluation of synthetic speech using the pesq measure", in *European Congress on Acoustics*, 2005.

[20] A. Rix, J. Beerends, M. Hollier en A. Hekstra, „Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs", in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, deel 2, 2001.

[21] M. Wester, C. Valentini-Botinhao en G. E. Henter, „Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations", in *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 09-2015, pp. 3476–3480.

[22] Fluency, *De stem van Arno*, 05-2014. adres: www.fluency.nl/stem/arno.htm.

[23] A. Group, *Voices List*, 2011. adres: http://www.acapela-vaas.com/ReleasedDocumentation/voices_list.php.

[24] Ivona, *Ivona SDK Technical specifications*, 2016. adres: http://4cwsuq3or162bqn011ym22w1.wpengine.netdna-cdn.com/wp-content/uploads/2016/11/IVONA-SDK-Technical-Specification-Sheet.pdf.

[25] L. Cordano, *Speech technologies in transport*, 09-2016.

[26] N. I. of Standards, T. (U.S.) en I. SEMATECH., *NIST/SEMATECH Engineering Statistics Handbook*. 2002. adres: https://books.google.nl/books?id=v-XXjwEACAAJ.

[27] A. G. Bluman, *Elementary statistics:* 7th ed. New York, NY: McGraw-Hill, 2009.

[28] B. A. Cerny en H. F. Kaiser, „A Study Of A Measure Of Sampling Adequacy For Factor-Analytic Correlation Matrices", *Multivariate Behavioral Research*, jrg. 12, nr. 1, pp. 43–47, 1977. adres: https://doi.org/10.1207/s15327906mbr1201_3.

[29] S. Tobias en J. E. Carlson, „Brief Report: Bartlett's test of sphericity and chance findings in factor analysis", jrg. 4, pp. 375–377, 07-1969.

[30] K. Chu, S. Dean en B. Illowsky, *Elementary Statistics*. Rice university Press, 2013, p.106 - 108.

[31] S. King en V. Karaiskos, „The Blizzard Challenge 2016", *The Centre for Speech Technology Research University of Edinburgh*, 2016.

[32] C. C. Arjan van Hessen Henk van den Heuvel, *Taal- en Spraaktechnologie voor visueel gehandicapten*, 2017. adres: https://notas.nl/artikelen/blogs/taal-en-spraaktechnologie-voor-visueel-gehandicapten.

[33] S. Cheh, *Japan's robot priest is ready to conduct and livestream your funeral*, 2017. adres: https://asiancorrespondent.com/2017/08/japans-robot-priest-ready-conduct-livestream-funeral/#MLdFhsOtzLLoxh6V.97.